On the Impact of Transposon Activity on Genome Evolution

Dissertation

zur

Erlangung der naturwissenschatlichen Doktorwürde

(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Stefan Roffler

von

Grüsch GR

Promotionskomitee

Prof. Dr. Beat Keller (Vorsitz)
PD Dr. Thomas Wicker (Leitung der Dissertation)
Prof. Dr. Christian von Mehring

Zürich, 2016

Table of contents

Summary

Transposable elements are mobile elements that have the ability to move and replicate within a genome. TEs are a major fraction of most plant genomes and were the main subject of my studies.

First, we investigated the activity of DNA transposons in rice. We compared the Asian rice with the closely related African rice. We analyzed polymorphisms related to the activity of DNA transposons on a genome-wide scale and found that DNA transposons are highly active in rice. Moreover, we found that the ratio between insertions and excisions significantly differs from the expected one for some TE superfamilies. This indicates that excisions of elements of some superfamiles might induce more drastic (and probably deleterious) re-arrangements than others. The excision process as such was often found to introduce deletions and “filler” DNA at the respective loci.

In a second project, we described a *Helitron* family (*DHH_Mothra)* which we found to be one of the most prominent DNA transposon in rice. Here, we could show, how transposons evolve from formerly autonomous elements to non-autonomous elements that lost all coding sequence. This form of extreme parasitism seems to be a very successful strategy among all transposons. Additionally, we showed that plant *Helitrons* have acquired a further protein most likely by horizontal transfer.

Moreover, we describe the influence of DNA transposon activity on genes and regulatory regions of grasses. We could show that DNA transposons preferably insert close by genes. The repair of the resulting double-strand break following excisions is error-prone and effects several kb around the excision site. By this they introduce an elevated mutation rate in the regulatory regions of genes and even alter the coding sequence significantly. This suggests DNA transposons to be a major force in the evolution of grass genomes.

In a side project, I was involved in the study of plant-pathogen-interactions, which encompassed several aspects on the plant and the pathogen side. Here, I established a method, which helped to identify an avirulence gene in wheat powdery mildew, the bulk segregant analysis.

Zusammenfassung

Transposable Elemente sind mobile Einheiten, die sich innerhalb eines Genoms bewegen und replizieren können. TEs machen einen Grossteil pflanzlicher Genome aus und waren Gegenstand meiner Arbeit.

Zuerst untersuchten wir die Aktivität von DNA Transposons in Reis. Dazu verglichen wir den asiatischen Reis mit dem nahe verwandten afrikanischen Reis. Wir analysierten TE-Polymorphismen im ganzen Genom und fanden, dass diese in Reis sehr aktiv waren. Zudem fanden wir, dass das Verhältnis von Insertionen zu Excisions (das Ausschneiden von TEs) für manche Superfamilien stark variieren. Dies deutet darauf hin, dass Excisions von manchen Transposon-Superfamilien vermehrt zu umfangreichen Re-Arrangements (und wahrscheinlich auch Deletionen) führen, als solche von anderen. Zudem fanden wir, dass nach Excisions häufig zusätzliche “Füller” DNA eingeführt wurde.

Als zweites Projekt beschrieben wir eine *Helitron* Familie (*DHH_Mothra*), die eines der meist verbreiteten DNA Transposon in Reis ist. Hier konnten wir zeigen, wie Transposons aus vorher autonomen Elementen zu nicht-autonomen Elementen werden, die jegliche codierende Sequenz verloren haben. Diese Form des Extrem-Parasitismus scheint eine erfolgreiche Strategie für alle Arten von Transposons zu sein. Darüber hinaus konnten wir zeigen, dass Pflanzen *Helitrons* ein zusätzliches Protein aufgenommen haben, höchst wahrscheinlich durch horizontalen Transfer.

Zusätzlich beschreiben wir den Einfluss von DNA Transposon Aktivität auf die regulative Elemente und Gene von Gräsern. Wir konnten zeigen, dass DNA Transposons vorzugsweise nahe bei Genen inserieren. Wenn sich TEs ausschneiden, resultiert daraus ein Doppel-Strang Bruch der DNA, dessen Reparatur fehlerbehaftet ist. Dies führt zu einer erhöhten Mutationsrate in den regulativen Regionen und selbst der Protein codierenden Bereiche der Gene. Dies legt nahe, dass DNA Transposons eine der Haupt-Antriebskräfte in der Evolution von Gras Genomen ist.

In einem Nebenprojekt habe ich mich mit Pflanzen-Pathogen-Interaktionen beschäftigt. Hier möchte ich die Bulk Segregant Analyse vorstellen, die mit dazu beigetragen hat, ein Avirulenz-Gen im Weizen Mehltau zu identifizieren.

Chapter 1:

General Introduction

The field of genomics is very versatile. With the ability to sequence Deoxyribose Nucleic Acid (DNA), many tools and methods for specific applications have been developed. Many involve extensive computational processing and have the power to address a wide range of biological questions. In the following chapter, I would like to introduce history, methods and technologies in the field of genomics. Furthermore, I provide background on TEs and their role in evolution and introduce the organisms I worked with during my PhD work.

1.1 Historical background on genomics

In contrast to many peoples believes, it was not the American biologist James Watson and the English physicist Francis Crick, who first discovered DNA in the late 1950ies (Pray, 2008). In fact, it was already identified in 1869 by a Swiss chemist, Friedrich Miescher, who first introduced the term “nuclein” (Dahm, 2005). Following Mieschers work, the Russian biochemist Phoebus Levene described a DNA to be composed of three units (phosphate, sugar (backbone) and base) and chemically resolved the structure of both RNA and DNA (ribose or deoxyribose as a sugar component, respectively) already in 1919 (Levene, 1919). Furthermore, the Austrian biochemist Erwin Chargaff developed a new paper chromatography method to separate and identify small amounts of organic material which are in principle still used these days and in 1950 he came up with two major conclusions: First, he found that the nucleotide composition of DNA varies among species, and second, he formulated "Chargaff's rule”, which says that the total amount of purines (A + G) equals total amount of pyrimidines (C + T). Inspired by x-ray pictures of Maurice Wilkins and Rosalind Franklin, it was finally Watson and Crick who were the first scientists that formulated an accurate description of the complementary, double-helical structure of DNA in 1953 (Watson and Crick, 1953). With the help of discussions with George Gamow, it was also Crick, who demonstrated in 1961 that three bases of DNA code of one amino acid in the Crick, Brenner, Barnett, Watts-Tobin experiment of 1961 (Crick *et al.*, 1961). The genetic code and its redundancy was thereon resolved by synthesizing of poly nucleotides by different scientists.

1.2. Technologies and Methods of DNA Sequencing

Sanger DNA Sequencing

It was Frederick Sanger in 1977, who developed a method to effectively sequence DNA that was thereafter most widely used for approximately 25 years and, for particular applications, still is today (Sanger *et al.*, 1977). The “Sanger” method makes use of di-deoxynucleosidetriphosphates (ddNTPs) that terminate strand elongation upon integration during *in vitro* DNA replication. Because ddNTPs lack a 3'-OH group, they can not form a phosphodiester bond to the next nucleotide like the deoxynucleosidetriphosphates (dNTPs) do and therefore stop DNA synthesis. The method moreover requires single-stranded DNA as template, random, small, radioactively labeled DNA primers and a DNA polymerase. To determine the sequence of a particular fragment, four separate reactions, one for each nucleotide, are needed. In each reaction, a small amount (approximately 1%) of the ddNTPs is added to the normal mix of dNTPs. As the ddNTPs are incorporated, the reaction stops at the location with a base complementary to the corresponding ddNTP that was added. The order of the nucleotides in the sequence can thus be inferred based on the fragment sizes using gel-electrophoresis. Each of the four the products for each ddNTP reaction are applied on a different lane. Since they all refer to the same radioactively labeled primers, they can be directly compared. Thus there will be one band for each position a respective dNTP was incorporated for each of the lanes. The results from the gel-electrophoresis are then transferred to a polymer sheet which is then exposed to x-ray autoradiography to be interpreted. While sequencing the 5.386 base pair genome of the bacteriophage φX174 was revolutionary in 1977, this was only the beginning. The modern version of Sanger sequencing, is much more time- and work-efficient. The four nucleotides are differently labeled with fluorescent dyes and can be applied as a mix. The DNA fragments then migrate through a capillary system to be automatically scanned by a laser.

To access the gene-containing portions of the genome, a technique that isolates mRNA and reversely transcribes it into its complementary DNA (cDNA) was developed. This method uses a poly A primer in a PCR-based procedure

with a reverse transcriptase (RT). In this way, the transcripts that were expressed at the timepoint of extraction are unspecifically amplified and can thereafter be sequenced. As a disadvantage, cDNA libraries lack the untranslated 5' and 3' regions and introns because they base on the finally spliced mRNA. The sequenced cDNAs are there often referred to as expressed sequence tags (ESTs).

Next Generation Sequencing (NGS)

In recent years, new methods have been established that produce manifolds the amount of sequence data in just a fraction of the time and cost. In 2005, the *Roche/454* pyrosequencing method was introduced and provided a platform which was capable of sequencing 25 million bases, at 99% or better accuracy in one four-hour run, an approximately 100-fold increase in throughput compared to the Sanger method (Marquiles *et al.*, 2005). The main advantages of this method is that DNA libraries are clonally amplified *in vitro* and then sequenced in parallel as a bulk, which enables the high throughput.

In the Roche/454 method, random DNA fragments are generated by shearing entire genomes into small fragments. The fragments are ligated to adapters which then bind to small beads that are single dispersed into manufactured, picolitre-sized wells. The wells are then “washed” in an emulsion based method in so called flow chambers. The four nucleotides bond to inorganic pyrophosphate are sequentially applied to the well. The release of the pyrophosphate upon the incorporation of the nucleotides leads to the emission of light that can be detected by an optical sensor from the bottom of each well. The method, however, is prone to errors, particularly in large homopolymers (seven or more), which can lead to ambiguous base calls. The lengths of the reads ranged from 100 to 400 bps when the system was introduced but has been further developed to exceed up to 700 bps by today (compared to approximately 900 – 1.000 bps for the Sanger method).

Currently, most widely used sequencing technology, however, is Illumina sequencing. In 2008, a method was introduced that allowed even higher throughput. Entire genomes could be sequenced at a high coverage at very low cost. Even though Illumina reads are relatively short compared to the previous

technologies (usually 100 to 300 bp), it could be used to determine polymorphisms between a reference to a very low cost. This so-called “genome re-sequencing” makes this technology even applicable for large scale diagnostics or screenings. Whereas the costs per megabase decreased approximately five-fold when *Roche/454* sequencing was introduced, it is now approximately 350.000 times cheaper than just 15 years ago (Figure 1).

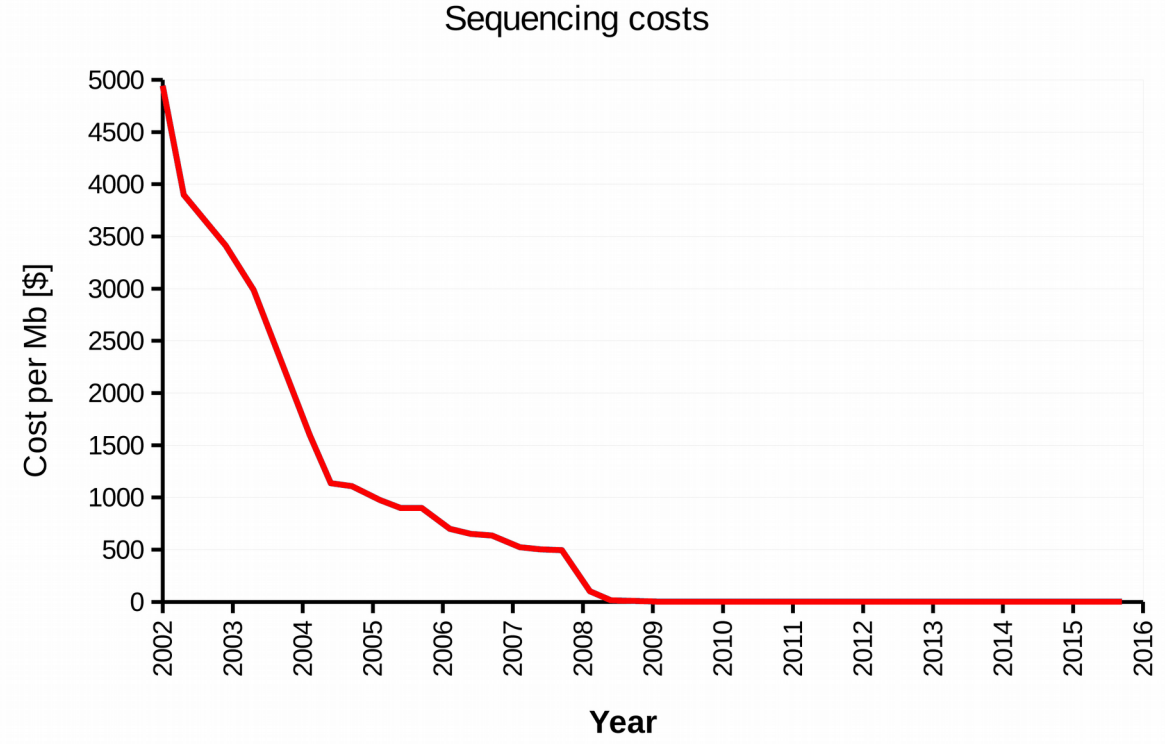


Figure 1. Course of the sequencing costs per megabase over the last 15 years (Source: http://www.genome.gov/sequencingcosts/).

Similar to *Roche/454* sequencing, Illumina uses sheared DNA fragments that are ligated to adapters. These are loaded onto a specialized chip where hundreds of thousands of oligonucleotides are anchored and bind to the DNA fragments. To generate templates, these fragments are thereafter amplified to approximately a thousand copies, in a phase called cluster generation. For the actual sequencing reaction, nucleotides with reversible 3' blockers are used. These force the primers to add only one nucleotide at a time. A mix of all four nucleotides is applied for each synthesizing step. Because the nucleotides have fluorescent tags, their wavelengths are specific for each nucleotide and are monitored for each individual spot of the chip by a camera. After washing non-incorporated nucleotides away, the 3’ terminal blocking group can be

chemically removed by a dye in a single step, allowing the start of the next cycle. Unlike pyrosequencing, a mixture of nucleotides can be applied so that all the DNA chains are extended one nucleotide at a time, making it more efficient. Moreover, the signal detection and image acquisition, allow for much larger arrays to be captured by sequential images taken from a single camera.
There are other technologies on the market such as SOLiD, Ion Torrent or the Single molecule real time (SMRT) sequencing of Pacific Biosciences (PacBio) that are still used and have their niches. In particular PacBio, that provides very long reads of up to 20 kbp (approximately 5 kbp on average) has also become more and more competitive. In particular for *de novo* sequencing projects of bacterial genomes or highly repetitive regions, this is of great value (see below). Here however, I focused on those technologies that are most frequently used and were also used for the projects described below. An overview of the technical hallmarks for each technology can be found in Table 1.

Table 1. Overview of the hallmarks for the most used sequencing technologies.

Method	**Read length**	**Cost per Mb**	**Strengh/weakness**
Sanger Sequencing	900 – 1,000 bp	~2,400 USD	long reads/ good quality
Roche/454	~ 700 bp	~ 10 USD	long reads / relatively fast / problems with homo-polymeres
Illumina	100 – 300 bp	~ 0,07 USD	high throughput / cheap
PacBio	several kb	~ 3,00 USD	very long reads / high error-rate

Bacterial Artificial Chromosome (BAC) - libraries

A major concept to also approach larger genomes is to reduce their complexity at different levels. One such technique, that has been widely used to create many reference genomes, is the use of bacterial artificial chromosomes (BACs) (O'Connor *et al.*, 1989). BACs are plasmids that contain large fragments of the genome of interest that was before digested by restriction enzymes. The fragment sizes depend on the restriction enzyme and genome but are usually 100 – 150 kbp in length. The fragments can then be cloned and amplified in *E. coli to* provide sufficient template. In this way, the sequencing and assembly of each BAC are reduced to a degree that could be accomplished. Similar to the assembly of individual reads, the BACs themselves can then be arranged into BAC contigs based on overlaps between one and another. This is done by "fingerprinting" the BACs with restriction enzymes. The individual patterns of the resulting fragment sizes are thereafter analyzed by the FPC (Soderlund *et al.*, 1997) or LTC (Frenkel *et al.*, 2010) softwares which determine overlaps and produce contigs. For sequencing, a minimum tiling path to cover the whole genome is selected. The respective BAC clones are then sequenced one by one. This "BAC-by-BAC" strategy is a rather conservative approach and is still used for the highest quality standards (Figure 2).

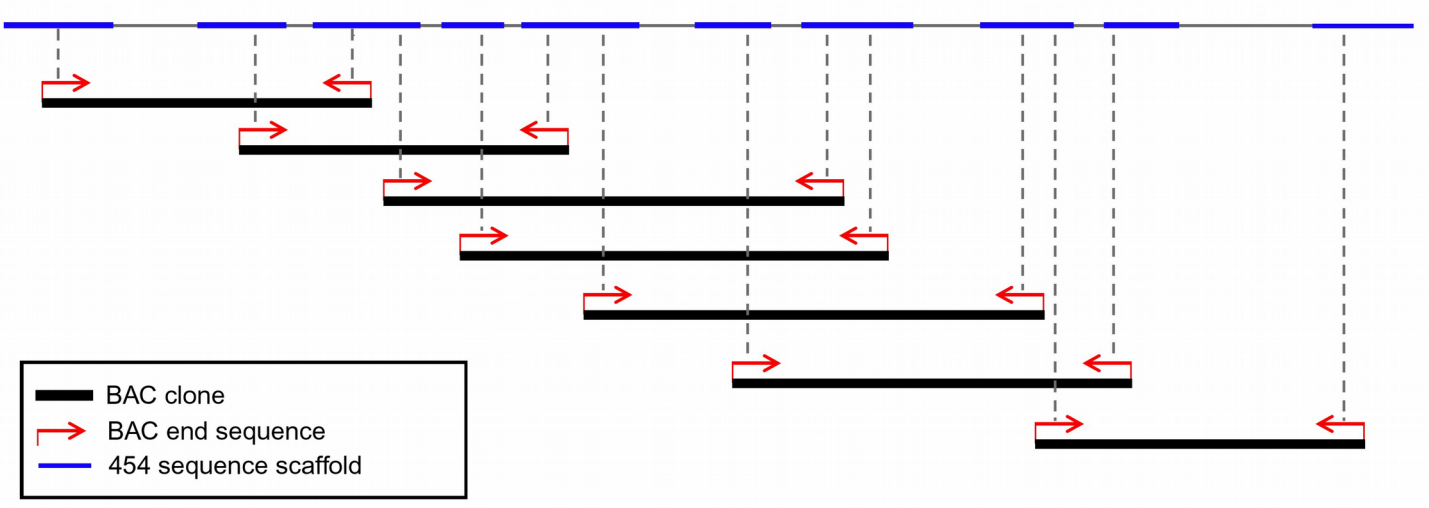


Figure 2. Schematic Illustration of the BAC-by-BAC approach. BAC clones are first assembled into BAC contigs. The BAC ends are sequenced from both ends to anchor the scaffolds produced by next generation sequencing (here Roche/454 scaffolds).

Paired end sequencing

Another major concept to close gaps between sequence contigs was to use paired end information (Edwards and Caskey, 1991). When generating a sequencing library, the DNA fragments are ligated to adapters or into vectors so that they can then be sequenced from both sides in opposite directions. Because the size range of the DNA fragments is known, it is possible to estimate the distance between both reads. This information can then be used to link two different sequence contigs that incorporate reads originating from the same pair into scaffolds. In 1995 Roach *et al.* introduced the use of fragments of varying sizes and proposed pure pairwise end-sequencing as possible strategy on large target genomes (Roach *et al.*, 1995). Accordingly, after sequencing simple microbial genomes, the first eukaryotic genomes (yeast in 1996) and the first multicellular species (*C. elegans* in 1998), the first entire plant genome of *Arabidopsis thaliana* (The Arabidopsis Genome Initiative, 2000) and finally the human genome (International Human Genome Sequencing Consortium, 2001) were sequenced using the paired-end Sanger technology in combination with the generation of BAC libraries (Schatz *et al.*, 2012). Other available plant reference genomes of such high quality include rice (*Oryza sativa)* in 2002, papaya (*Carica papaya)* in 2008 and maize (*Zea mays)* in 2009 (Schatz *et al.*, 2012). Because of its high accuracy and the relatively long individual reads of up to 1.000 bp, Sanger sequencing is still used in experiments with small sample size (e.g. to verify cloning products).

1.3. Bioinformatics

To handle the increasing amount of data that could be created with Sanger sequencing, computer assisted tools were needed already early on. An approach that is still known as the shotgun method, was first proposed in 1979 by Staden (Staden, 1979). His method describes the fragmentation of a genome into small, random pieces that are then individually sequenced. The innovation of this approach was that a computer program that would assemble the reads based on overlaps between individual fragment to create continuous DNA fragments (so called contigs), circumventing the additional step of creating genetic and BAC-based maps. Because of limited computational resources at this time, this was still limited to small genomes such as bacteriophages. With the increasing performance of the sequencing techniques, bioinformatic tools needed to evolve accordingly. Even though the prices and availability of computational power have developed similar to that of genomic sequencing, it is currently still one of the biggest challenges to analyze all the data that is generated.

Genome assembly

Whereas older approaches put great emphasis on the optimal exploitation of all reads and accounted for issues like the correction of sequencing errors, they were incapable of assembling very high numbers of reads, even on very large computers. The first software to assemble the large quantities of 454 data, was the software Newbler, that was released by the developers of the 454 method at *Roche*. For the Illumina data, on the other hand, several independent research groups and commercial companies have developed assemblers that are all based on similar principles. The first one was released was SHARCGS (Dohm *et al.*, 2007) which was quickly followed by diverse others such as CLC cell, ABySS, SOAP or Velvet (reviewed by Miller *et al.*, 2010). They have different strengths and weaknesses but all use heuristic approaches such as eliminating overrepresented reads or indexing them. The de Bruijn graph approach, which is used by CLC Genomics Workbench, split the reads into even smaller fragments, so called k-mers, that are indexed and incorporated in a network frame to be processed more efficiently (Compeau *et al.*, 2011). Even

though Illumina sequencing was originally designed for re-sequencing rather than for *de novo* assemblies, such algorithms opened up the possibility to use this technology also for *de novo* sequencing.
Another approach is to combine different datasets. For example for creating the *B. graminis f.sp. tritici* genome (Wicker *et al.,* 2013), which was sequenced in our group, the contigs assembled from 454 reads could be anchored to ends of BAC clones that were sequenced with the Sanger method. Finally, Illumina reads where mapped on the assembly to correct for sequencing errors (see below, Figure 2).

Gene annotation

After having sequenced and assembled genomes, scientists need to cope with this large mount of data. Today, there are dozens or even hundreds of specified programs that are usually streamlined into genome annotation pipelines. These differ in details, but share common features (reviewed by Yandell and Ence, 2012):
In bioinformatics, there are two basic approaches to identify genes on genomic sequences. The first is based the homology to known genes. Programs such as BLAST (Altschul *et al.*, 1990) provide functions that search all possible reading frames of an input nucleotide sequence for possible protein sequences that are in the database or *vice versa*. The methods have been further developed such as for example the Position-Specific Iterative BLAST (PSI-BLAST) search, which is supposed to identify more distantly related sequences by iteratively creating some consensus sequence of the best hits that would then be re-used for another round of search (Altschul *et al.*, 1997). Because this approach is all based on homology to existing sequences, It is of great importance to congregate this information in public databases such as GenBank at NCBI, the EMBL Data Library or SWISS-PROT and of course the awesome TREP database (www.botinst.uzh.ch/en/research/genetics/thomasWicker/TREP.html). Thus the main drawback of the homology-based approach is its relying on existing databases.
The second approach to gene annotation is to predict genes *ab initio.* In bacteria, which lack introns, such programs would simply predict every open

reading frame as a potential gene, which can be problematic in eukaryotes. It is almost impossible to accurately identify intron-exon boundaries of genes. In maize, for example, intron sizes can extend as much as 70 kb (Hurni *et al.*, 2015). Moreover, eukaryotic genomes often carry pseudo-genes or have different splicing variants. To predict intron-exon boundaries of eukaryotic genes, there are programs that therefore make use of transcriptomic data such as ESTs or RNA-seq data, if available. Most modern gene prediction pipelines such as MAKER, have combined all approaches on homology, on expression and *de novo* to obtain the best results (Cantarel *et al.*, 2008).

From genomes to genes of interest

Based on the large amounts of data that is available, it is possible to create networks that link phenotypes with genetic variants in genome-wide association studies (GWAS) (Hirschhorn and Mark, 2005 / Haines *et al.*, 2005). GWAS studies are based on information on polymorphisms between species, varieties or even individuals that is generated by mapping the sequence reads of one species to a reference. The nucleotides that differ between the two studied organisms, so called single nucleotide polymorphisms (SNPs), are thought to determine the differences between the two. Performing multiple such comparisons, those SNPs that are associated with certain phenotypes are candidates for being causative for the phenotype. However, this method very much relies on the quality of the reference genome, the degree of homology between the genomes and a consistent phenotyping. Moreover, many traits are controlled by multiple genes that interact in complex ways, leading to noisy signals which makes it hard to identify true positives.

Phylogenetic Analysis

Phylogenetics analysis are widely used in evolutionary biology to infer the evolutionary history and relationship of genes. Every phylogenetic analysis is based on an alignment of two or more genomic sequences. Each alignment is then scored according to a probability matrix that accounts for different base substitution models. The sequences being likely to belong to one group are positioned next to each other in a phylogenetic tree. Modern programs such as Mr Bayes (Altekar *et al.*, 2004) use complex mathematical procedure such as Monaco-Markov-Chains that simulate trees using all known substitution-matrices to estimate the most likely constellation which would lead to the observed data.

Here, it is important to notice that genes and gene families within genomes do not necessarily evolve at the same pace. For example house-keeping genes, that are essential for survival, underlie strong selection and thus evolve slower than for example a virulence gene that constantly has to adapt (see below). For a conclusive phylogenetic analysis, careful selection of genes is therefore important. *Parra et al.* (Parra *et al.*, 2006) determined a set of 458 core proteins that are well conserved in a wide range of eukaryotes. These genes thus are well suited to study distantly related species but are less informative when comparing organisms that diverged only recently. Genes, specific for a species or just a group of organisms, have often originated from more recent gene duplication or recombination events (see below). Therefore, it is important that true orthologs are compared, which means those genes (or gene copies) that have the same origin in a common ancestor and not more recent copies that may only exist in one but not the other organism.

1.4. Repetitive DNA and the “C-value-paradox”

One of the major challenges in genomics are repetitive sequences. They severely affect almost every step of an analysis, starting from sequencing, to assemblies, mappings, annotations and thus the methods based on it (such as GWAS). It was already known in the 1970ies that the genomes of different eukaryotes are very variable in size by estimating nuclear DNA amounts through densitometric measurements (Bennett and Smith, 1976). Indeed, the genome size of *Amoeba dubia* is approximately estimated to reach 670 gigabases (Gregory, 2001) whereas the microsporidium *Encephalitozoon cuniculi* comprises only 2.9 Mbp (Biderre *et al*., 1995, Katinka *et al.*, 2001). Interestingly, the number of genes is more or less constant for most eukaryotes, an observation known as *“C-value-paradox”* (Thomas, 1971).

It has been elucidated, that the genome sizes of larger monophyletic groups are of comparable sizes, especially for animals and fungi. For example, genomes of mammals, that diverged between 70 and 113 million years ago (MYA), all comprise about 3,000 Mbp (Gheerbrandt *et al.*, 2005), whereas reptile and bird genomes, which evolved about 240 MYA (Benton, 1993), average at around 1,000 Mbp in size (Krishan *et al.*, 2005). In contrast, the so far known fungal genomes are all comparably small. They range from few Mbp, such as the 12 Mbp genome of yeast (Mackiewicz *et al.*, 2002) to approximately 180 Mbp such as the wheat powdery mildew genome (Wicker *et al.*, 2013). It is surprising how these genome sizes remained so similar, given that the main evolutionary lineages of fungi have diverged more than 700 MYA (Taylor and Barbee, 2006).

In contrast to animals and fungi, plant genomes vary much more even between closer related species. The approximately 120 Mbp size the of *Arabidopsis thaliana* genome, for example, is one of the smallest so far described (The Arabidopsis Genome initiative, 2000). The closely related *Brassica* species, however, have genomes that are 5 – 10 times larger, even though they only diverged 15-20 MYA (Yang *et al.*, 1999). Within monocotyledonous plants, these differences are even more extensive. The grasses *Brachypodium dystachion* (The International Brachypodium Initiative, 2010) and rice (The International Rice Genome Sequencing Project, 2005) have genome sizes of 273 and 389

Mbp, respectively. This is considerably larger than the *Arabidopsis* genome but substantially smaller than the genomes of some agriculturally important grasses such as wheat (The International Wheat Genome Sequencing Consortium, 2014) or barley (The International Barley Sequencing Consortium, 2012) with estimated haploid genome sizes of approximately 5,700 Mbp.

With advances in sequencing, it became more and more clear that the numbers of protein coding genes are very similar to each other. Whereas there were first 100,000 genes predicted for the human genome (International Human Genome Sequencing Consortium, 2001), this number has been reduced to approximately 30,000, similar to other vertebrates like mice (Mouse Genome Sequencing Consortium, 2002) or chicken (International Chicken Genome Sequencing Consortium, 2004). This number also corresponds to the expected gene number in rice, where initially 60,000 genes were proposed. The more recent annotation of the *Brachypodium distachyon* genome, that used very stringent parameters, revealed 25,554 genes (The International Brachypodium Initiative, 2010), which is very similar to that of the most recent version of the *Arabidopsis thaliana* genome where 25,498 genes were annotated (The Arabidopsis Genome Initiative, 2000). In contrast, fungi and invertebrate animals have approximately half or even less genes. Yeast, with its compact 12 Mbp genome has less than 6,000 genes (Mackiewicz *et al.*, 2002) while insects such as *Drosophila melanogaster* have approximately 13,000 genes (Adams et al., 2000). The immense differences between genome sizes thus can not be explained by the number of genes.

The Transposable Element Invasion

As more and more genomic sequences became available, it soon became clear that most of the sequences that determine genome size are in fact derived from transposable elements (TEs) which in some cases can be found in thousands of copies. Transposons are small genetic units that can replicate themselves or move around in the genome, or both (see below). Because they often code for only few or even no proteins, they make use of the host reproduction machinery. Therefore, they have often been termed "junk", "selfish" or “parasitic” DNA (Orgel and Crick, 1980). TEs can have sizes of

several kb and often contain large arrays of very repetitive nucleotide patterns. Moreover, they often insert into one and another to form complex, nested arrangements. Because TEs can have thousands of copies of a particular TE family in some organisms (usually those with large genomes such as wheat or barley), they are also overrepresented in the sequence reads, resulting in tremendous amounts of ambiguous fragments. Many assembly algorithms can not clearly assign such reads and usually discard them, which terminates the elongation of sequence contigs. When assembling a large and repetitive genome such as the wheat genome, this results in hundreds of thousands of contigs of rather short length. Sequencing technologies that produce longer reads, such as the Sanger, *Roche/454* or PacBio technologies, thus can deal better with this issue than Illumina. To sequence and assemble most of the more unique, gene coding regions of a genome, however, also short reads are usually sufficient. Shatalina *et al.*, for example, constructed a high resolution genetic map based Illumina sequences to identify a quantitative trait locus for resistance to glume blotch in wheat. Here, the chromosomes 3B of two wheat lines were flow-sorted and sequenced with Illumina. The assembly of the sequence contained thousands of genes that provided the basis to design the SNPs for the genetic map (Shatalina *et al.*, 2014).

When *de novo* annotating a genome, transposons are usually masked at a very early stage. Because TEs can carry gene fragments and, if the element is autonomous, possess their “own” genes, this can lead to false positive gene annotations. The helicase and the RPA-homolog of the *DHH_Mothra* Helitron familiy, which is described later (Chapter 3, Roffler *et al.*, 2015), would be such examples. However, to mask repetitive sequences, they must first be identified. An accurate annotation and classification for such elements is therefore of great importance. The sequences used for masking are usually consensus sequences derived from multiple copies of similar elements.

TE Classification

In 2007, Wicker *et al.* (Wicker *et al.*, 2007) proposed a TE classification system which included the definition of consistent criteria that are characteristic features of the main TE superfamilies and families (Figure 3). Moreover, a three-letter-code based naming system was proposed which would identify the TE on each of the following levels.

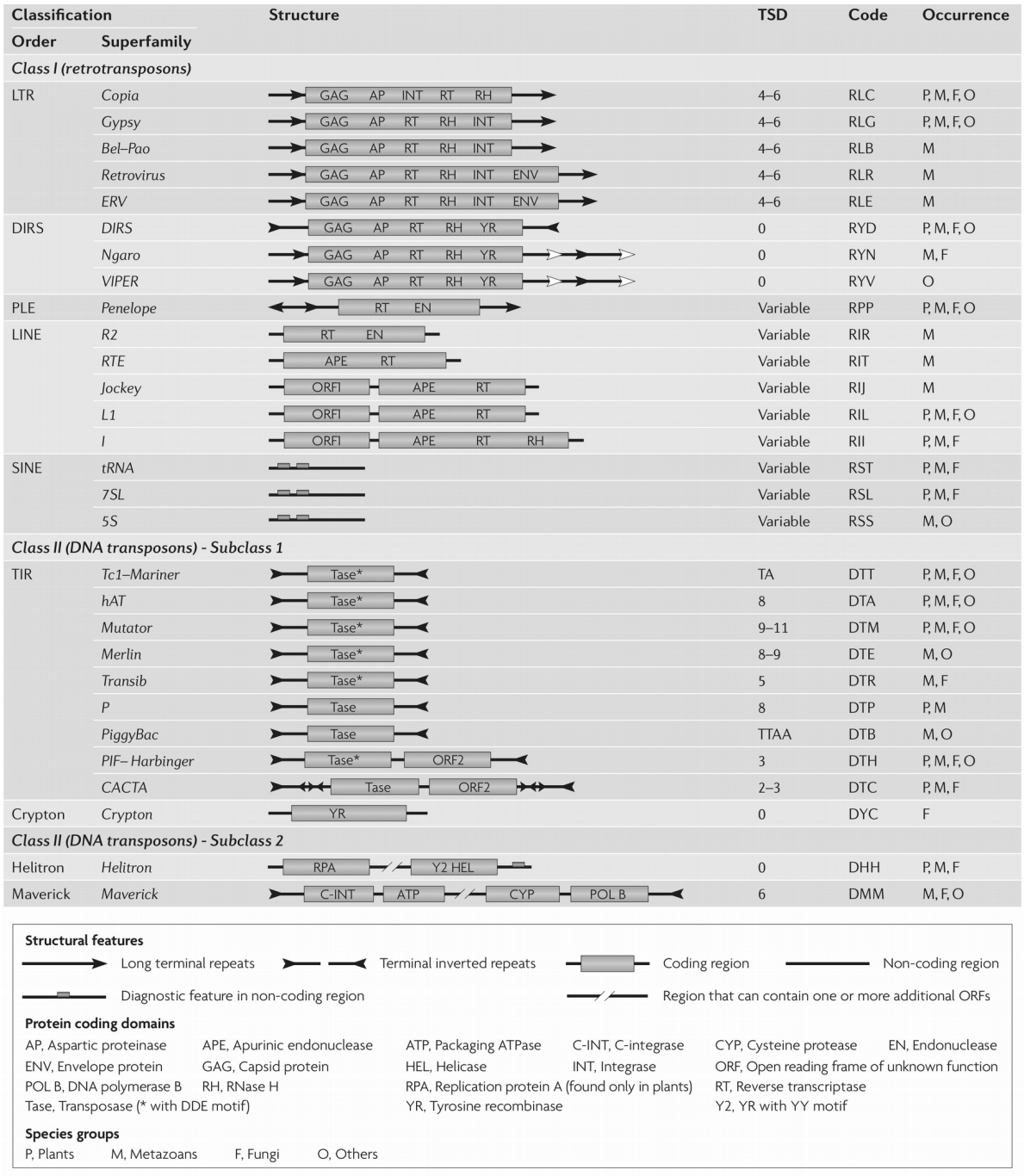


Figure 3. The proposed TE classification system by Wicker *et al.* (2007) (adapted from Wicker *et al.* 2007).

First of all, transposons are grouped into two major classes, which are further subdivided into 9 orders and 29 superfamilies. Class 1 TEs or “retrotransposons” contain all TEs which replicate via an mRNA intermediate in a "copy-and-paste" process. Class 2 elements, are DNA transposons, that move their DNA itself analogous to a "cut-and-paste" process. Here we will emphasize more on Class 1 TEs because retrotransposons of the Long Terminal Repeat (LTR) order are the most abundant type of transposons in plants and retrotransposons of the LINE and SINE order are the most prominent elements in wheat powdery mildew. DNA transposons were the subject of the three publications that are part of this thesis and explained in detail in the respective chapters (see below).

Going back to the *C-value-paradoxon*, it was shown that bursts of LTR-retrotransposon activity lead to the dramatic enlargement of many plant genomes (El Baidouri and Panaud, 2013). In maize, for example, they make up to 50 % of the genome (SanMiguel *et al.*, 1998). Because of their “copy and paste” life-cycle, these genomes were flooded with additional copies, sometimes of only few TE families. The wild relative of rice, *Oryza australiensis,* for example, almost doubled its genome size in the last three million years through acquisition of more than 90,000 LTR-retrotransposon copies of mainly three TE families (Piegu *et al.*, 2006).

LTR-retrotransposons can range from few hundreds baise pairs up to 25 kb in size (Neumann *et al.*, 2003). Autonomous elements have the capacity to encode a GAG capsid protein, a reverse transcriptase and a integrase protein. The size of their characteristic LTRs range from a few hundred base pairs to more than 5 kb. LTRs have a conserved “TG” at their start and a “CA” at their 3' end. Another characteristic feature of LTR retrotransposons are target site duplications (TSDs) of 4–6 bp which are generated upon insertion. Probably the best described example for a LTR retrotransposon is BARE1, that belongs to the *Copia* superfamily. BARE1 has reached approximately 16,000 copies in the barley genome and is still highly active in barley (Vicient *et al*., 1999).

Interestingly, retroviruses, even though they have long been described as viruses (Wicker *et al.*, 2007), are closely related to LTR retrotransposons. They share the same structure of domains and might thus most likely have evolved

from Gypsy LTR retrotransposons that acquired a set of additional proteins such as an envelope protein (ENV) (Frankel and Young, 1998) and additional regulatory sequences (Seelamgari 2004).

Whereas the LTR retrotransposons are the predominant retro-elements in plants, members of the long interspersed nuclear element (LINE) order seem to be more successful in vertebrates and fungi. In the wheat powdery mildew, LINEs are the most abundant retrotransposons (Parlange *et al.*, 2011). Also in mammals, the LINE L1 superfamily is one of the most abundant, numbering about $8,5^{5}$ copies or about 21% of the human genome (Boissinot *et al.*, 2000). LINEs are divided in five superfamilies, lack LTRs and can reach several kilobases in length. The superfamilies are based on the protein domains that they contain. They all contain a reverse transcriptase ORF. LINEs of the R2 superfamily, however, contain an endonuclease ORF but lack a packing ATPase which can be found in the other four superfamilies (RTE, Jockey, L1 and I). The elements of the Jockey, L1 and I superfamilies moreover contain an additional ORF1 of unknown function (Wicker *et al.*, 2007).

Autonomous and non-autonomous TEs

One key feature, that might have contributed to the great success of TEs, is that they are able to lose their coding regions but retain the ability to proliferate by captivating the enzymes of related TEs or the host itself. This phenomenon was observed for almost all superfamilies, in retrotransposons as well as in DNA transposons (explained in more detail below in Roffler and Wicker, 2015 and Roffler *et al.*, 2015). Elements of the short interspersed element (SINE) order, for example, lack coding sequence and are therefore usually only a few hundred base pairs in size. They depend on trans-acting functions such as the RT from LINE elements (Kramerov and Vassetzky, 2005 / Dewannieux et al., 2003 / Kajikawa and Okada, 2002). In contrast to the described non autonomous DNA transposons (see below) SINEs are not deletion derivatives of autonomous class I elements. They originate from accidental retrotransposition of various polymerase III (Pol III) transcripts that possess an internal Pol III promoter, allowing them to be expressed (Kramerov and Vassetzky, 2005). Similar to LINEs, SINEs are also very prominent in powdery

mildew. Even though they are only from two families, they occupy approximately 3 % of the genome despite their small sizes (Parlange *et al.*, 2011). In the human genome, the approximately 1,5 million SINE elements of the *Alu* family (13% of the genome) double the number of LINE elements (Deininger, 2011).

TEs and genome evolution

Transposon activity has heavily influenced the evolution of eukaryotic genomes. Bursts of transposon activity have been associated with the radiation and diversification of many animals such as the primate (Pace and Feschotte, 2007) or bat lineages (Prittham and Feschotte, 2007). The influence of LTR retrotransposons on the genome sizes of many plants is undisputed (see above). TE can moreover influence gene expression. Insertions of TEs can lead to loss of function phenotypes. Most prominent is the discovery of Barbara McClintock who found that an insertion of a DNA transposon leads to differently pigmented maize corns (McClintock, 1953). The LTR regions of LTR retrotransposons, moreover, contain promotors which can lead to activation or differential expression of genes (Klaver and Berkhout, 1994). Another general mechanism that is more relevant on an evolutionary timescale is that Class II TEs preferably insert close to or into regulatory regions of genes. Thereby, TE activity also increases the mutation rate of several kb flanking regions of the site, influencing even the coding parts of genes (see Wicker and Roffler, 2016 (submitted)). Other processes that TEs contributed to are exon shuffling (Jiang *et al.*, 2004; Lai *et al.*,2005; Morgante *et al.*, 2005; Paterson *et al.*, 2009) or gene movement in general (Wicker *et al.*,2010). Additionally, it has been shown that specific TE families are essential for centromere and telomere function in some species (Wolfgruber *et al.*, 2009; Frydrychova *et al.*, 2008). Also the V(D)J-recombination system of the mammalian immune system was derived from TE recombinases (Jones, 2004). Taken together, there are many aspects of genome structure and evolution of central functions that have been influenced by TE mechanisms that make TE a major factor for evolution in general.

1.5. Organisms studied in this thesis

The rice (*Oryza*) genus

Rice is, with wheat and maize, among the top three nutrient supplying plants world-wide (Food and Agriculture Organization of the United Nations, http://www.fao.org/home/en/). Compared to both other crops, its genome size is relatively small. For these reasons, it was one of the first fully sequenced plant genomes. In 2002, the International Rice Genome Sequencing Project (IRGSP) released a draft version of a BAC-by-BAC Sanger sequenced genome, covering approximately 93 % of the genomic space (Goff *et al.*, 2002). A first map-based sequence that provided full chromosomes, virtually all of the euchromatin and two complete centromeres was published in 2005 (International Rice Genome Sequencing Project, 2005). The genome and its annotation, however, are constantly improved. In the analysis that we performed (Roffler and Wicker, 2015a and Roffler *et al.* 2015) we used the fifth version of this genome as our reference sequence. The most recent release (version 7) reports a total of 373 Mb of non-overlapping sequence from the 12 rice chromosomes, covering almost the entire genome which is expected to have a size between 384.2 and 386.5 Mb (Kawahara *et al.*, 2013). In total 55,986 genes were annotated. Noteworthy, these include 16,941 gene loci that are derived from TE sequences.

The genus *Oryza* has evolved within the last 15 million years and contains 23 species that are organized in ten genome types (Kim *et al.*, 2008). As part of an ongoing project to separately sequence and assemble ten *Oryza* genome types at best possible quality, the genome of *Oryza glaberrima*, the African rice, was published in 2014 (Wang *et al.*, 2014). The TE content of *O. glaberrima* is expected to be lower (104 Mb) than the one of *O. sativa* (156 Mb) which would account approximately for the difference in overall genome size between the two (Wang *et al.*, 2014). Even though, both species have been domesticated for more than 3,000 years, they clearly are of different origin. Their estimated divergence time is approximately 600,000 years (Ammiraju *et al.*, 2008). However, specific accessions of *O. sativa* and *O. glaberrima* are sexually compatible. Traits such as drought tolerance or pathogen resistance thus make

O. glaberrima attractive for breeders (Sarla and Swamy, 2005).
Taking together the expected high TE activity, the relatively recent divergence time of the species and the fact that both genomes were independently sequenced and assembled with the highest standard make these two genomes an excellent system to study TE activity (see Roffler and Wicker, 2015, Roffler *et al.,* 2015 and Wicker *et al.,* 2016 (submitted)).

The powdery mildews

Wheat, is aside of maize and rice, one of the top three nutrient suppliers for mankind with a production of more than 713 million tonnes in 2013 (Food and Agriculture Organization of the United Nations, www.fao.org/home/en/). Because its extensive cultivation (usually in mono-cultures) it is an attractive target for many pathogens such as fungi, bacteria, viruses or nematodes. The wheat powdery mildew (*Blumeria graminis f.sp. tritici*) is an obligate biotrophic Ascomycete and one of the most devastating wheat pathogen world-wide and can cause tremendous yield losses (Glawe, 2008). Powdery mildew is among the widest spread fungal pathogens and infects almost 10,000 angiosperm species (reviewed by Glawe, 2008).
Powdery mildews are ascomyetes of the order *Erysiphales.* The genus *Blumeria,* which includes powdery mildews that grow on grasses, is divided into eight *formae specialis (f.sp.)*. For example, *Blumeria graminis f.sp. tritici (B.g. tritici)* is the powdery mildew that grows on wheat. The specificity of individual isolates can even be race-specific. This means that a certain *B.g. tritici* isolate will only grow on some wheat cultivars while other wheat cultivars will recognize the pathogen and block the fungal attack. Mildew evolution, however, is not limited to strict co-evolution with their host (Troch *et al.*, 2014). Modern approaches involving NGS sequencing of numerous isolates revealed that hybridization between closely related *formae speciales* can lead to host-range expansions (Menardo *et al.*, 2016).

The Blumeria life-cycle

Blumeria has a sexual and an asexual life-cycle. In the asexual life-cycle, a conidiospore lands on the leaf of a potential host and produces a primary germ

tube. This is believed to sense the leaf and prime the target cell for an upcoming attack (Edwards, 2002). In the next step, a secondary germ tube emerges which forms an appressorium at its tip. It is not entirely clear which mechanisms are used to achieve a successful penetration of the cuticula. It is possible that chemical or enzymatic factors are involved but most likely mechanical force plays a major role. The penetration peg which emerges from the appressorium is thought to create high local pressures (Both *et al.*, 2005). Once the cell is invaded, the fungus establishes a haustorium inside the cell. This fungal feeding organ almost completely invaginates the host cell, but it does not penetrate the plasma-membrane. Once the haustorium is established, the fungus will start to grow secondary hyphae that also attack neighboring cells and allow the fungus to colonize the whole leaf. After successful colonization of a leaf, the nutrients will be directed into reproduction. Hundreds of thousands of clonal spores are produced on top of the leaf which appear to the human eye as white powder. This “powder” is then dispersed by wind to infect surrounding plants and fields.

Sexual reproduction is believed to occur at most once every year (Wicker *et al.* 2013). At the end of the season, when the environmental conditions get harsher, two individuals of opposite mating-type can fuse and form cleistothecia. These structures are much more resistant to abiotic stresses and make it suitable to overcome winter. Importantly, this is the only phase in which genetic material between different isolates can be recombined. Overall, the main reproduction mode is asexual (Wicker *et al.*, 2013).

The obligate biotroph lifestyle shapes the genome

The reference genome for the wheat powdery mildew, the Swiss isolate 96224 has been sequenced by our group and was published in 2013 (Wicker *et al.*, 2013). The approach combined a classic BAC library with the whole genome shotgun method using the *Roche*/454 technology. The BAC clones were restriction finger printed and arranged into 250 finger print contigs using the FPC software (Soderlund *et al.*, 1997). The cumulative size of the BAC contigs was 180 Mb. The ends of all BAC clones were moreover sequenced from both sides using the Sanger method. The contigs resulting from the *Roche*/454 data

could thereafter be anchored to the FPC contigs via the BAC-end-sequences (Figure 2). Sequence gaps that could be estimated on the basis of the FP contigs were filled with strings of Ns (Wicker *et al.*, 2013). The final assembly consists of 126 Mb whereof 107 Mb could be anchored to the FP contigs. The number of non-N bases comprises 82 Mb whereof 67 Mb could be anchored to FP contigs. To correct for sequencing errors (proneness for polynucleotides of the Roche/454 technology), Illumina sequences where mapped on to the final contigs.

More than 90% of the genome was classified as TE sequences. Despite the large genome size, only 6,540 genes were identified. However, 98% of the CEGMA eukaryotic core genes were found in full length, indicating that gene space was covered nearly completely. Many gene families from the primary and secondary metabolism were absent or only partially present (Wicker *et al.*, 2013). This can be explained by the obligate biotrophic lifestyle of *Blumeria*. The haploid nature of the *Bgt* genome and the clonal propagation into millions of descendants create an environment with extreme intraspecific competition, which leads to population dynamics similar to that in bacteria. The loss of genes can thereby be a strategy to reduce cost for reproduction. Wicker *et al.* (2013) showed that even essential metabolic pathways such as synthesis of certain amino acids were lost in *B.g. tritici.* It is believed that they are elicited from the host.

Importantly, among the roughly 6,000 genes, there are more than 700 candidate effector genes. This substantial portion of the gene content is probably due to the obligate biotrophic lifestyle. Effectors are believed to be needed to invade the host and/or to manipulate and reprogram cellular processes. Therefore, the mildews require a broad and highly specialized palette of effectors (see below in Chapter 2).

1.6. Overview and aims of projects covered in this thesis

The main overarching aim of my PhD work was to study the role of transposon activity in genome evolution. My work was divided into several projects of which four are covered in this thesis.

The aim of the first project was to survey and study polymorphic TEs in the two rice species *O. sativa* and *O. glaberrima*. We were interested in studying the types of sequence rearrangements that are caused by transposon insertions and excisions. The results of this study were published in *Mobile DNA* (Roffler and Wicker, 2015). Additionally, the annotated DNA transposons became part of the RiTE database (Copetti *et al.,* 2015).

The aim of the second project was to describe the *DHH_Mothra* family*,* a novel family of *Helitron* DNA transposons ubiquitous in rice. We identified several putatively autonomous and semi-autonomous derivatives of which some contain an additional protein exclusive to plant *Helitrons.* We could show that this protein was most likely acquired by horizontal transfer. The results of this study were published in *Mobile DNA* (Roffler *et al.*, 2015).

The aim of the third project was to show the impact of DNA transposon activity on genes. Based on the findings in rice, we could show that in particular excisions and the resulting DSB repair introduce a significantly higher mutation rate in the untranslated but also the coding regions of genes. This mechanism could moreover be demonstrated in maize, wheat and barley, indicating error-prone DNA repair as a major evolutionary force on the genes of grasses. The results of this study are ready for submission (Wicker *et al., in preparation*).

As a side project I was involved in genomics of powdery mildew. The main subject of our research group is the study of plant pathogen interactions. This involved the assembly and annotation of the draft genome of powdery mildew, the development of a tool for bulk segregant analysis to identify avirulence loci in the haploid genome of powdery mildew (Chapter 2), the assistance with SNP analysis, effector identification, locus annotation, mappings, alignments and other general bioinformatics support. The results contributed to several publications (Wicker *et al.,* 2013, Bourras *et al.,* 2015, Parlange *et al.,* 2015, Menardo *et al.,* 2016).

1.7. List of publications to which this PhD work contributed

Wicker T, Yu Y, Haberer G, Mayer KFX, Reddy Marri P, Rounsley S, Chen M, Zuccolo A, Panaud O, Wing RA, Roffler S: **DNA transposons specifically accelerate evolution of genes in rice and other grasses.** *In preparation*.

Menardo F, Praz CR, Wyder S, Ben-David R, Bourras S, Matsumae H, McNally KE, Parlange F, Riba A, Roffler S, Schaefer LK, Shimizu KK, Valenti L, Zbinden H, Wicker T, Keller B: **Hybridization of powdery mildew strains gives rise to pathogens on novel agricultural crop species.** *Nature Genetics 2016,* **48:**201-205.

Roffler S, Menardo F, Wicker T: **The making of a genomic parasite – the *Mothra* family sheds light on the evolution of *Helitrons* in plants.** *Mobile DNA 2015,* **6:**23.

Roffler S and Wicker T: **Genome-wide comparison of Asian and African rice reveals high recent activity of DNA transposons.** *Mobile DNA 2015,* **6:**8.

Copetti D, Zhang J, El Baidouri M, Gao D, Wang J, Barghini E, Cossu RM, Angelova A, Maldonado L CE, Roffler S, Ohyanagi H, Wicker T, Fan C, Zuccolo A, Chen M, Costa de Oliveira A, Han B, Henry R, Hsing YI, Kurata N, Wang W, Jackson SA, Panaud O, Wing RA: **RiTE database: a resource database for genus-wide rice genomics and evolutionary biology.** *BMC Genomics 2015,* **16:**538.

Bourras S, McNally KE, Ben-David R, Parlange F, Roffler S, Praz CR, Oberhaensli S, Menardo F, Stirnweis D, Frenkel Z, Schaefer LK, Flückiger S, Treier G, Herren G, Korol AB, Wicker T, Keller B: **Multiple avirulence loci and allele-specific effector recognition control the *Pm3* race-specific resistance of wheat to powdery mildew.** *Plant Cell 2015,* **27:** 2991-3012.

Parlange F, Roffler S, Menardo F, Ben-David R, Bourras S, McNally KE, Oberhaensli S, Stirnweis D, Buchmann G, Wicker T, Keller B: **Genetic and molecular characterization of a locus involved in avirulence of Blumeria graminis f. sp. tritici on wheat Pm3 resistance alleles.** *Fungal Genet Biol. 2015,* **82:**181-192.

Wicker T, Oberhaensli S, Parlange F, Buchmann JP, Shatalina M, Roffler S, Ben-David R, Dolezel J, Simkova H, Schulze-Lefert P, Spanu PD, Bruggmann R, Amselem J, Quesneville H, Ver Loren van Themaat E, Paape T, Shimizu KK, Keller B: **The wheat powdery mildew genome shows the unique evolution of an obligate biotroph.** *Nature Genetics 2013,* **45(9):**1092-98.

1.8. References:

Adams MD *et al*.: **The Genome Sequence of *Drosophila melanogaster.*** *Science 2000,* **287(5461):**2185-2195.

Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F: **Parallel Metropolis-coupled Markov chain Monte Carlo for Bayesian phylogenetic inference.** *Bioinformatics 2004*, **20:**407-415.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J. Mol. Biol. 1990,* **215:**403-410.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucl. Acids Res. 1997,* **25(17):**3389-3402.

Ammiraju JSS et al.: **Dynamic evolution of Oryza genomes is revealed by comparative genomic analysis of a genus-wide vertical data set.** *Plant Cell 2008*, **20:**3191–3209.

Bennett MD and Smith JB: **Nuclear DNA amounts in angiosperms.** *Philos Trans R Soc Lond B Biol Sci 1976*, **274:**227–274.

Bentley DR et al.: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature 2008,* **456:**53-59.

Benton MJ: **Reptilia**. Pp. 681-715 in Benton MJ, ed. **The fossil record 2.** *Chapman & Hall, London (1993)*.

Biderre C, Pages M, Metenier G, Canning EU, Vivaras CP: **Evidence for the smallest nuclear genome (2.9 Mb) in the microsporidium *Encephalitozoon cuniculi*.** *Mol Biochem Parasitol 1995*, **74:**229–231.

Boissinot S, Chevret P, Furano AV: **L1 (LINE-1) Retrotransposon Evolution and Amplification in Recent Human History.** *Mol. Biol. Evol. 2000,* **17(6):**915-928.

Both M, Csukai M, Stumpf MPH, Spanu PD: **Gene Expression Profiles of *Blumeria graminis* Indicate Dynamic Changes to Primary Metabolism during Development of an Obligate Biotrophic Pathogen.** *Plant Cell 2005,* **17(7):**2107-2122.

Cantarel BL, Korf I, Robb SMC *et al*.: **MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes.** *Genome Research 2008,* **18(1):**188-196.

Compeau PEC, Pevzner PA, Tesler G: **How to apply de Bruijn graphs to genome assembly.** *Nature Biotechnology 2011,* **29:**987-991.

Crick FH et al. Barnett L, Brenner S, Watts-Tobin RJ: **General nature of the genetic code for proteins.** *Nature 1961,* **192(4809):** 1227–32.

Dahm R: **Friedrich Miescher and the discovery of DNA.** *ScienceDirect 2005,* **278(2):**274-288.

Deininger P: ***Alu* elements: know the SINEs.** *Genome Biology 2011,* **12:**236.

Dewannieux M, Esnault C, Heidmann T: **LINE-mediated retrotransposition of marked Alu**

sequences. *Nature Genet. 2003,* **35:**41–48.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing.** *Genome Research 2007,* **17(11):**1697–706.

Edwards A and Caskey T: **Closure strategies for random DNA sequencing.** *Methods: A Companion to Methods in Enzymology 1991,* **3(1):**41–47.

Edwards HH: **Development of primary germ tubes by conidia of *Blumeria graminis f.sp. Hordei on leaf epidermal cells of Hordeum vulgare.*** *Canadian Journal of Botany 2002,* **80(10):**1121-1125.

El Baidouri M, Panaud O: **Comparative Genomic Paleontology across Plant Kingdom *Reveals the Dynamics of TE-Driven Genome Evolution.*** *Genome Biol. Evol. 2013,* **5(5):**954-965.

Frankel AD and Young JA: **HIV-1: fifteen proteins and an RNA.** *Ann. Rev. Biochem. 1998,* **67:**1–25.

Frenkel Z, Paux E, Mester D, Feuillet C, Korol A: **LTC: a novel algorithm to improve the efficiency of contig assembly for physical mapping in complex genomes.** *BMC Bioinformatics 2010*, **11:**584.

Frydrychova RC, Biessmann H, Mason JM: **Regulation of telomere length in Drosophila.** *Cytogenet Genome Res 2008*, **122:**356–364.

Gheerbrandt E, Domning DP, Tassy P: **Paenungulata (Sirenia,Proboscidea, Hyracoidea, and relatives)**. Pp. 84-105 in Rose KD, and Archibald JD, eds. ***The rise of placental mammals: origins and relationships of the major extant clades*.** *Johns Hopkins University Press*, Baltimore (2005).

Glawe DA: The Powdery Mildews: A Review of the World's Most Familiar (Yet Poorly Known) Plant Pathogens. *Annual Review of Phytopathology 2008,* **46:**27-51.

Goff SA, Ricke D, Lan TH, Presting G, Wang R et al.: **A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. *Japoinca*).** *Science 2002,* **296:**92-100.

Gregory TR: Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. *Biol Rev Camb Philos Soc 2001*, 76:65–101.

Haines JL, Hauser MA, Schmidt S, Scott WK *et al*.: **Complement Factor H Variant Increases the Risk of Age-Related Macular Degeneration.** *Science 2005,* **308(5720):**419-421.

Hirschhorn JN and Mark JD : **Genome-wide association studies for common diseases and complex traits.** *Nature Reviews Genetics 2005,* **6:**95-108.

Hurni S, Scheuermann D, Krattinger SG, Kessel B, Wicker T, Herren G, Fitze MN, Breen J, Presterl T, Ouzunova M, Keller B: **The maize disease resistance gene *Htn1* against northern corn leaf blight encodes a wall-associated receptor-like kinase.** *PNAS 2015.* **112(28):**8780-8785.

International Chicken Genome Sequencing Consortium: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.** *Nature 2004,* **432(9):**695-716.

International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature 2001,* **409:**860-921.

International Rice Genome Sequencing Project. **The map-based sequence of the rice genome.** *Nature 2005*, **436:**793–800.

Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR: **Pack-MULE transposable elements mediate gene evolution in plants.** *Nature 2004*, **431:**569–573.

Jones JM and Gellert M: **The taming of a transposon: V(D)J recombination and the immune system.** *Immunol. Rev. 2004,* **200:**233-48.

Kajikawa M and Okada N: **LINEs mobilize SINEs in the eel through a shared 3' sequence.** *Cell 2002*, **111:**433–444.

Katinka MD, Duprat S, Cornillot E, Metenier G *et al.*: **Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi.** *Nature 2001*, **414:**450-453.

Kawahara Y *et al*.: **Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data.** *Rice 2013*, **6:**4.

Kim H, Hurwitz B, Yu Y, Collura K, Gill N, SanMiguel P, Mullikin JC *et al.*: **Construction, alignment and analysis of twelve framework physical maps that represent the ten genome types of the genus Oryza.** *Genome Biology 2008,* **9(2):**R45.

Klaver B and Berkhout B: **Comparison of 5' and 3' long terminal repeat promoter function in human immunodeficiency virus.** *Journal of Virology 1994,* **68(6):**3830-3840.

Kramerov D and Vassetzky N: **Short retroposons in eukaryotic genomes.** *Int. Rev. Cytol. 2005,* **247:**165–221.

Krishan A, Dandekar P, Nathan N, Hamelik R, Miller C, Shaw J: **DNA index, genome size, and electronic nuclear volume of vertebrates from the Miami Metro Zoo.** *Cytometry A 2005*, **65:**26–34.

Lai J, Li Y, Messing J, Dooner HK: **Gene movement by *Helitron* transposons contributes to the haplotype variability of maize.** *PNAS 2005*, **102:**9068–9073.

Levene PA: **The Structure of Yeast Nucleic Acid: IV. Ammonia Hydrolysis.** *J. Biol. Chem. 1919.* **40:**415-424.

Mackiewicz P *et al*.: **How many protein-coding genes are there in the Saccharomyces cerevisiae genome?** *Yeast 2002,* **19(7):**619-629.

Marquiles M, Egholm M, Altman WE *et al.*: **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature 2005,* **437(7057):**376-80.

McClintock B: Induction of Instability at Selected Loci in Maize. *Genetics 1953,* **38(6):**579-99.

Miller JR, Koren S, Sutton G: **Assembly algorithms for next-generation sequencing data.** *Genomics 2010,* **95(6):**315-327.

Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A: **Gene duplication and exon shuffling by *Helitron*-like transposons generate intraspecies diversity in maize.** *Nature Genetics 2005*, **37:**997–1002.

Mouse Genome Sequencing Consortium: **Initial sequencing and comparative analysis of the mouse genome.** *Nature 2002,* **420:**520-562.

Neumann P, Pozarkova D, Macas J: **Highly abundant pea LTR retrotransposon Ogre is constitutively transcribed and partially spliced.** *Plant Mol. Biol. 2003,* **53:**399–410.

O'Connor M, Pfeifer M, Bender W: **Construction of large DNA segments in Escherichia coli.** *Science 1989,* **224(4910):**1307-1312.

Orgel LE and Crick FHC: **Selfish DNA: the ultimate parasite.** *Nature 1980*, **284:**604-607.

Ouyang S, Zhu W, Hamilton J, Haining L, Campbell M, Childs K, et al.: **The TIGR Rice Genome Annotation Resource: improvements and new features.** *Nucleic Acids Res. 2007,* **35:**D883–7.

Pray L: **Discovery of DNA structure and function: Watson and Crick.** *Nature Education 2008*, **1(1):**100.

Pace II JK, Feschotte C: **The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage.** *Genome research 2007,* **17(4):**422-432.

Parlange F, Oberhaensli S, Breen J, Platzer M, Taudien S, Simkova H, Wicker T, Dolezel J, Keller B: **A major invasion of transposable elements accounts for the large size of the Blumeria graminis f.sp. tritici genome.** *Functional & Integrative Genomics 2011,* **11(4):**671-677.

Parra G, Bradnam K, Korf I: **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.** *Bioinformatics 2007,* **23(9):**1061-1067.

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, *et al*. **The *Sorghum bicolor* genome and the diversification of grasses.** *Nature 2009*, **457:**551–556.

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, et al.: **Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice.** *Genome Res. 2006,* **16:**1262–9.

Pritham EJ, Feschotte C: **Massive amplification of rolling-circle transposons in the lineage of the bat Myotis lucifugus.** *PNAS 2007,* **104(6):**1895-900.

Roach JC, Boysen C, Wang K, Hood L: **Pairwise end sequencing: a unified approach to genomic mapping and sequencing.** *Genomics 1995,* **26(2):**345-353.

Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors.** *Proc. Natl. Acad. Sci. U.S.A.* **74(12):**5463-7.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL: **The paleontology of intergene retrotransposons of maize.** *Nature Genetics 1998*, **20:**43–45.

Schatz MC, Witkowski J, McCombie WR: **Current challenges in *de novo* plant genome sequencing and assembly.** *Genome Biol. 2012,* **13:**243.

Sarla N and Swamy BM: **Oryza glaberrima: a source for the improvement of Oryza sativa.** *CURRENT SCIENCE-BANGALORE 2005*, **89(6)**:95.

Seelamgari A *et al.*: **Role of viral regulatory and accessory proteins in HIV-1 replication.** *Front. Biosci. 2004,* **9:**2388–2413.

Shatalina M, Messmer M, Feuillet C, Mascher F, Paux E, Choulet F, Wicker T, Keller B: **High-resoultion analysis of a QTL for resistance to Stagonospora nodorum glume blotch in wheat reveals presence of two distinct resistance loci in the target interval.** *Theor. Appl. Genet. 2014*, **127:**573–586.

Sim GK, Kafatos FC, Jones CW, Koehler MD, Efstratiadis A, Maniatis T: **Use of a cDNA library for studies on evolution and developmental expression of the chorion multigene families.** *Cell 1979, **18(4):**1303–16.*

Soderlund C, Longden I, Mott R: **FPC: a system for building contigs from restriction fingerprinted clones.** *Comput. Appl. Biosci. 1997*, **13:**523–535.

Staden R: **A strategy of DNA sequencing employing computer programs.** *Nucleic Acids Res. 1979,* **6(7):** 2601-2610.

Taylor JW and Berbee ML: **Dating divergences in the Fungal Tree of Life: review and new analyses.** *Mycologia 2006*, **98:**838–849.

The Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana.*** *Nature 2000,* **408:**796-815.

The International Barley Genome Sequencing Consortium: **A physical, genetic and functional sequence assembly of the barley genome.** *Nature 2012,* **491:**711-716.

The International Brachypodium Initiative: **Genome sequencing and analysis of the model grass *Brachypodium distachyon.*** *Nature 2010,* **463:**763-768.

The International Wheat Genome Sequencing Consortium (IWGSC): **A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome.** *Science 2014,* **345:**6194.

Thomas CA: **The genetic organization of chromosomes.** *Annu. Rev. Genet. 1971*, **5:**237–256.

Troch V, Audenaert K, Wyand RA, Haesaert G, Höfte M, Brown JKM: **Formae speciales of cereal powdery mildew: close or distant relatives?** *Molecular Plant Pathology 2014,* **15(3):**304-314.

Vicient CM, Kalendar R, Anamthawat-Jonsson K, Schulman AH: **Structure, functionality, and evolution of the BARE-1 retrotransposon of barley.** *Genetica 1999,* **107:**53-63.

Wang M *et al.*: **The genome sequence of African rice (Oryza glaberrima) and evidence for independent domestication.** *Nature Genetics 2014*, **46(9)**:982-988.

Watson JD and Crick FH: **Molecular Structure of Deoxypentose Nucleic Acids.** *Nature 1953*, **71:**737-738.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al.: **A unified classification system for eukaryotic transposable elements.** *Nat Rev Genet. 2007,* **8(12):**973–82.

Wicker T, Buchmann JP, Keller B: **Patching gaps in plant genomes results in gene movement and erosion of colinearity.** *Genome Res. 2010,* **20:**1229-1237.

Wickens MP, Buell GN, Schimke RT: **Synthesis of double-stranded DNA complementary to lysozyme, ovomucoid, and ovalbumin mRNAs. Optimization for full length second strand synthesis by Escherichia coli DNA polymerase I.** *J. Biol. Chem. 1978,* **253(7):**2483-95.

Wolfgruber TK, Sharma A, Schneider KL et al.: **Maize centromere structure and evolution: sequence analysis of centromeres 2 and 5 reveals dynamic Loci shaped primarily by retrotransposons.** *PLoS Genet 2009*, **5:**e1000743.

Yandell M and Ence D: **A beginner's guide to eukaryotic genome annotation.** *Nature Reviews Genetics 2012,* **13:**329-342.

Yang YW, Lai KN, Tai PY, Li WH: **Rates of nucleotide substitution in angiosperm mitochondrial DNA sequences and dates of divergence between *Brassica* and other angiosperm lineages.** *J Mol Evol 1999*, 48:597–604.

Chapter 2:

The *AvrPm3*-gene: Wanted dead or alive!

The main research field of our group is plant-pathogen-interactions. Therefore, I was involved in genomics of the wheat powdery mildew. One of the goals was to identify and close a avirulence gene. The results of this work became part of the publication “Multiple Avirulence Loci and Allele-Specific Effector Recognition Control the Pm3 Race-Specific Resistance of Wheat to Powdery Mildew” by Bourras *et al.*, published in *Plant Cell 2015.*

2.1. Introduction

The plant immune system differs from the mammalian one. Plants lack a systemic and complex adaptive immune system like the V(D)J recombination system of mammals and therefore rely much more on inherited immune defense. As opposed to animals, plants are not mobile, and each cell requires its own, fully functional immune system. To meet this end, the plants have evolved resistance genes (also called R-genes), and established a complex system that mainly relies on two pillars. The first one, also referred to as the basal defense system, is triggered by the recognition of conserved patterns that are shared and often piovital for wide groups of possible invaders. The recognition of chitin, which is a central component of fungal cell walls, or flagellin the principal substituent of the bacterial flagellum, are just two examples for these pathogen- or microbe-associated molecular patterns (PAMPs or MAMPs / reviewed by Jones and Dangl, 2006). Upon infection, these patterns are recognized by pattern recognition receptors (PRRs) which are mostly trans-membrane proteins that then activate various defense responses via kinase signal cascades. This PAMP triggered immunity (PTI) can be, for example, the deposition of callose to fortify the cell wall or the accumulation of phenols, reactive oxygen species and other chemical compounds (Luna *et al.*, 2011). Moreover, the cells respond with profound transcriptional changes to pathogen induced signals (Hückelhoven and Panstruga, 2011). The regulated genes are often shared by responses to abiotic stresses such as drought, wind or extreme temperatures and often involve hormones such as salicylic acid, jasmonic acid or ethylene (Denancé *et al.*, 2013).

Many pathogens found ways to suppress PTI using effectors. This effector triggered susceptibility (ETS) takes place inside the plant cell and involves small, secreted proteins. Effectors are relatively loosely defined (Spanu, 2012): They are approximately 150 amino acids in size and have an N-terminal signal peptide to be delivered to the host. Moreover, they share two conserved cysteine residues and an N-terminal Y/F/WxC-motif (Godfrey *et al.*, 2010). They could either directly interfere with the PAMP receptors or, once delivered into the host, with one of their downstream targets. The defense reaction is thereby suppressed. Effectors are also know as virulence factors or virulence genes. To

counter ETS, plants have evolved resistance genes that encode proteins which recognize effectors. This class of genes mostly encode for proteins that contain a nuclear binding site (NBS) and a leucine rich repeat (LRR) domain and are therefore called *NBS-LRR*-genes (reviewed by McHale *et al.*, 2006). There are different classes of these genes. Some have an amino-terminal Toll/Interleukin 1 receptor homology region (TIR) whereas other have coiled-coiled domains. If an effector protein is recognized by the plant, this triggers a “hypersensitive” response (HR) which leads to cell death to prevent the pathogen from growing. What formerly was a virulence factor thus turns into an avirulence gene *(Avr).* Even though the mechanism is based on interactions between proteins, it is generally described as “gene-for-gene” interaction (Flor, 1971). This permanent “genetic arms race” creates strong selection pressures for both sides. The pathogen has to evolve and/or acquire new effectors as well as lose or modify those that became avirulence genes. The host in turn must recognize new effectors and modify its own proteins so they can not be targeted by effectors anymore.

In 2004, Yahiaoui *et al.* cloned the *NBS-LRR* gene *Pm3* in wheat which confers resistance against powdery mildew to hexaploid wheat (Yahiaoui *et al.*, 2004). So far 17 functional, true alleles of *Pm3* have been identified and tested for their resistance spectra (Bhullar *et al.*, 2010). Interestingly, the resistance spectra of some *Pm3* alleles overlap. In particular, the spectra of the *Pm3f* and *Pm3c* alleles are completely covered by the recognition spectrum of *Pm3a* and *Pm3b*, respectively. Therefore, the alleles *Pm3a* and *Pm3b* are considered “stronger” alleles than *Pm3f* and *Pm3c* (Brunner *et al.*, 2010).

Functional studies indicated that slight sequence alterations of the leucine rich repeat (LRR) domain are responsible for differences in allele specificity (Brunner *et al.*, 2010) which suggests a direct interaction of the resistance protein with the respective variants of an AVR effector. Moreover, the particular high sequence identity of >97% among all *Pm3* alleles indicates recent diversification of the *Pm3* allelic series (Yahiaoui *et al.*, 2009).

Taken together, the recent evolution of the *Pm3* alleles and the (presumably) direct interaction between *AvrPm3* (the effector gene in *Bgt*) and *Pm3* (the resistance gene in wheat) provides a promising system to study the genetic

and molecular mechanisms controlling specificity on both sides, the host as well as the pathogen. However, in this context, identification of the *Avr* partners of the *Pm3* alleles is a necessity.

Because powdery mildew has a haploid genome, every allele is “dominant” which would always lead to a 1 : 1 segregation ratio following the Mendelian law. In a previous study, Parlange et al. crossed the *Bgt* isolates 96224 (avirulent) and JIW2 (virulent) (Parlange *et al.*, 2015) and identified a locus, presumably controlling avirulence towards Pm3C and Pm3F. The locus was genetically mapped to a genomic interval of 26 kb which segregated almost perfectly in a 1:1 ratio. On this locus 1, a single putative effector gene was identified as the best candidate gene (*BCG1*). Even though the genetic linkage was very strong for this locus, *BCG1* could not be functionally proven to be the Avr, when transiently expressed in wheat containing the respective resistance gene (*Pm3C* and *Pm3F*). Moreover, the F1 population showed an additional, intermediate phenotype on *Pm3C* containing lines, whereas the phenotypes for *Pm3F* were clearly virulent or avirulent.

Here, I describe the development of a method to identify genetic regions in the *Bgt* genome that are responsible for avirulence on Pm3F containing wheat lines. The aim of this project was to identify effector candidate genes to further test for function. We combined classical genetics and modern sequencing technologies to in a genome-wide association study that helped identify candidate AvrPm3 genes. The results became part of the publication by Bourras *et al.* 2015 which was published in the *Plant Cell* (Bourras *et al.*, 2015).

2.2. Methods

Bulk segregant analysis (BSA)

To examine the inheritance of *Bgt* avirulence genes, the isolates 96224 and 94202 were crossed by Bourras and McNally (Bourras *et al.*, 2015) and phenotyped on wheat lines containing six different *Pm3* alleles. Because isolate 96224 is recognized by wheat containing either of the *Pm3a-Pm3f* alleles and isolate 94202 is virulent to all these alleles, they were considered promising candidates for a cross. The 167 F1 recombinant, haploid progeny were then raised on *Pm3* containing lines that would recognize the *AvrPm3* gene.

The population segregated in a 1:1 ratio on lines containing the alleles *Pm3a*, *Pm3c*, and *Pm3e* (Figure 1A). For the *Pm3F* allele, a ratio of 1:3 (A:V) progeny suggests two independent loci controlling the phenotype (Figure 1B). Interestingly, the ratio for the *Pm3F* allele segregated in a 1:1 ratio in the previous cross by Parlange (Parlange *et al.*, 2015). The ratios on the *Pm3B* (5:1:2) and *Pm3D* (2:1:5) containing lines, however, indicate a third locus which we will here not consider in detail (Table 1).

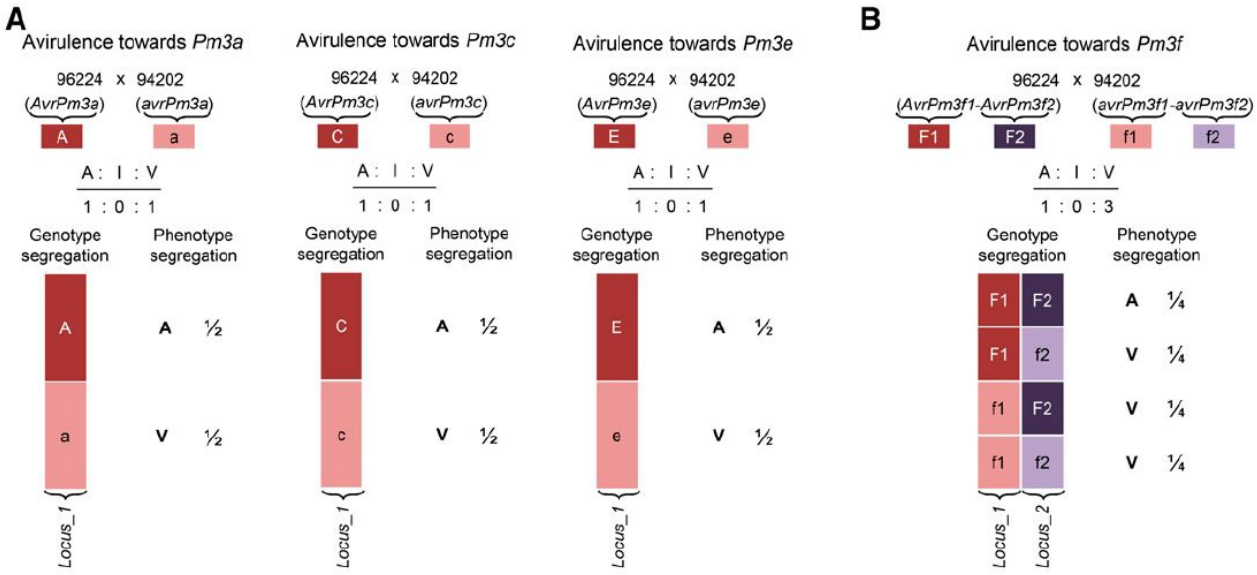


Figure 1. Overview of the segregation of the F1 progeny of the 96224 x 94202 cross phenotyped on *Pm3* alleles A, C, E and F. The 1:1 ratio indicates a single genetic locus to be responsible for avirulence for alleles Pm3a, Pm3c and Pm3e. Pm3f, however, shows a ratio of 1:3 (A:V), suggesting two independent loci to control avirulence (adapted from Bourras *et al.* 2015).

Table 1. Complete overview of F1 progeny segregating on different wheat lines containing race-specific avirulence genes (adapted from Bourras *et al.*, 2015).

Pm3 Resistance		Progeny Phenotype[a]			Genetic Segregation		
Alleles	Cultivar/Line[b]	A	I	V	Ratio[c]	χ^2	P[d]
Pm3a	Asosan^{8*CC}	81	0	85	1:0:1	0.096	0.756
Pm3b	Chul^{8*CC}	108	17	33	5:1:2	2.319	0.313
Pm3c	Sonora^{8*CC}	77	1	81	1:0:1	0.107	0.743
Pm3d	Kolibri	44	18	96	2:1:5	0.744	0.689
Pm3e	W150	35	1	32	1:0:1	0.147	0.701
Pm3f	Michigan Amber^{8*CC}	46	2	119	1:0:3	0.745	0.388

[a]Progeny phenotype was scored as LC. Phenotypes are indicated as "A" for avirulent (LC = 0), "I" for intermediate avirulent (LC = 10 to 40%), and "V" for virulent (LC = 60 to 100%).
[b]Wheat cultivars/lines carrying the *Pm3* allele indicated in the first column. Near-isogenic lines obtained after backcrossing eight times in cultivar 'Chancellor' are indicated by the superscript "8*CC."
[c]Genetic segregation is given as a theoretical ratio of A:I:V. Deviation of phenotype numbers from the theoretical ratios was tested with the χ^2 test for goodness of fit. The degree of freedom was assigned as the number of phenotypic classes – 1.
[d]The probability value for the χ^2 test. $P < 0.05$ would indicate significant deviation from the theoretical ratios.

First, markers from both populations, the 94202 cross by Bourras and McNally and the JIW2 cross by Parlange, were used to create a genetic map (Bourras *et al.,* 2015): Illumina sequences of both respective isolates were mapped on the reference sequence of the avirulent isolate 96224 for Kompetitive Allele-Specific PCR (KASP) technology analysis (He *et al.*, 2014). Selected SNPs were then tested on 164 F1 progeny from the JIW2 population and 154 from the 94202 population, which yielded in 251/254 and 224/228 polymorphic SNP loci, respectively. In combination with an additional 80 amplified fragment length polymorphism (AFLP) markers that were available for the JIW2 population (Parlange *et al.*, 2015), both populations produced maps with 17 linkage groups. Because 200 of the SNP markers were designed to be shared between the two populations, it was possible to produce three, yet fragmented, consensus linkage groups containing all the loci controlling the AvrPm3-Pm3 interactions. Consistent with the observation that a subset of progeny is avirulent on all Pm3 alleles, one locus was genetically interacting with all the six tested alleles (Locus 1). Locus 1 was the one also mapped before by Parlange *et al.* which provided a useful positive control for the BSA experiment. By selecting different subsets of progeny, a second locus interacting with the Pm3f allele (Locus 2) and a third locus interacting with the Pm3b, Pm3c, and Pm3d alleles (Locus 3) could be genetically identified.

High resolution mapping of *AvrPm3f2*

For the fine mapping of the AvrPm3f2 locus, a subset of 70 F1 progeny from the 94202 cross was used. These were all avirulent on Pm3A and Pm3C and segregated in a 1:1 ratio, thus indicating that they all have the avirulent genotype of the parent 96224 at Locus 1 (AvrPm3f1). Because this subset of progeny segregates in a 1:1 ratio (A:V) also on Pm3f, we could distinguish AvrPm3f2 and avrPm3f2 located on Locus 2 (Figure 2).

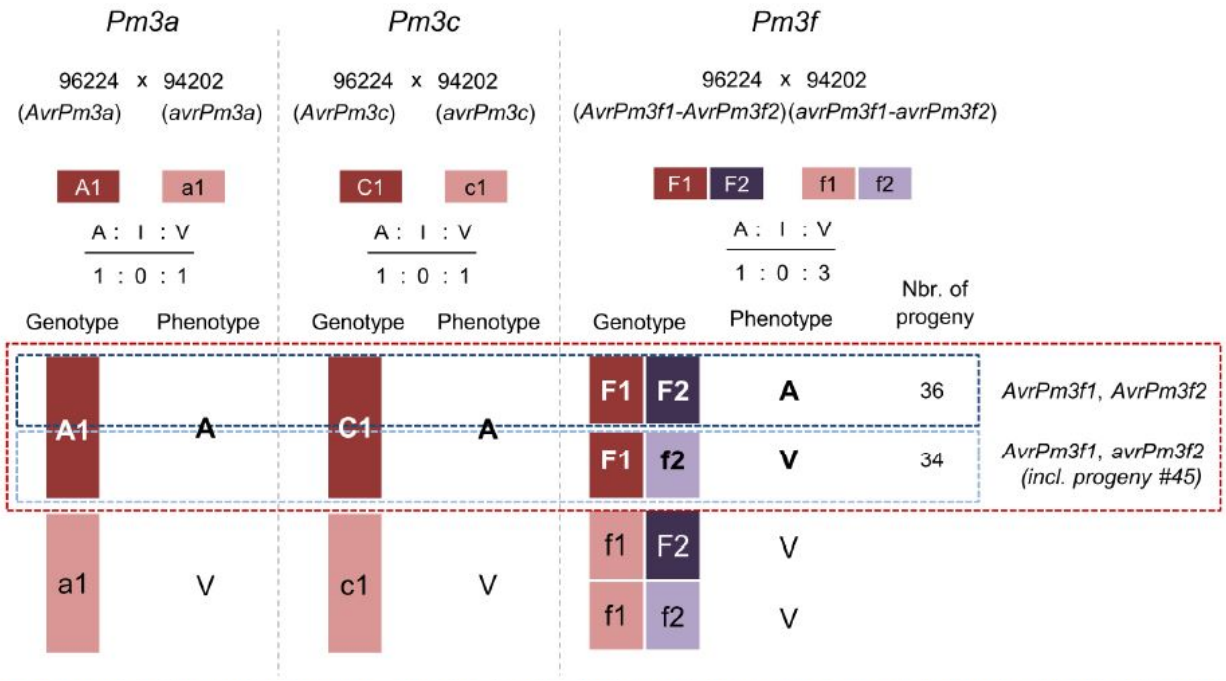


Figure 2. Subset of 70 F1 progeny that were used for the high resolution mapping of *AvrPm3f2.* Finally, 23 progeny of the AvrPmF1/AvrPmF2 genotype were selected for BSA (adapted from Bourras *et al.*, 2015).

NGS sequencing and assembly

For the bulk sequencing we used Illumina sequencing. Given the genome size of approximately 180 Mbp we generated 407 million reads of 100 bp length to aim an average coverage of approximately 200x-fold for the whole bulk (approximately nine for each of the 23 progeny). For the mapping we used CLC Genomincs Workbench (Version 5) on standard parameters. For filtering SNPs and visualization we used in house Perl scripts.

2.3. Results

As mentioned above, we will here focus on the interaction with the Pm3F allele and describe a method which helped to identify Locus 2. While the genetic interval of Locus 1 has been well described by Parlange *et al.* (2015), there were still many gaps and no co-segregating markers for Locus 2. Because of the enormous amount of repetitive sequences in *Bgt*, its assembly resulted in a very fragmented genome. The BSA helped to design additional markers for high resolution mapping which eventually helped to select the BAC clones and identify the gene *AvrPm3F-Pu7*. The KASP genotyping data from the selected 70 progeny was used to generate a low-resolution genetic map on which *AvrPm3f2* was flanked within an interval of 19 cM between the markers M033RE and M426MI (Table 2).

Bulk sequencing

To further reduce the genetic interval, we used whole-genome sequencing of F1 progeny in a process similar to the bulk segregant analysis described by Takagi *et al.* (2013). For our analysis, 23 progeny of the AvrPm3F1/AvrPm3F2 genotype (that carries the avirulent parent alleles at both loci (F1,F2)) were selected and Illumina sequenced in an equal ratio as one bulk. According to Mendel, we would expect a 1:1 distribution of both parental isolates for haploid genomes. This means for each polymorphic position in the bulk genome, 50 % of the mapped reads would come from each haplotype. However, for the regions we selected (that contain the AvrPm3F1/F2 locus) we expect 100 % of the reads to correspond to the respective phenotype. When mapping the bulk data on the reference genome, 265 million reads could be aligned, resulting in an average read depth of 168. On the raw dataset, a total of 336,309 SNPs were identified. Quality filtering with cutoffs of > 20x and < 400x coverage resulted in 135,210 SNPs. The threshold for the SNP calling was chosen to ensure representation of each progeny and to exclude SNPs in transposable elements based on the distribution of the coverage per SNP.

Table 2. Overview of the initial FPC-contigs that were selected by the BSA analysis (adapted from Bourras *et al.*, 2015).

FPC-contigs[1]	Linkage[2]	Marker ID[3]	Genetic distance[4]
Ctg-172	*AvrPm3f1*	M172RE	4.7 cM
Ctg-346	*AvrPm3f1*	M346MI	15.3 cM
Ctg-057	*AvrPm3f1*	M057RE	17.3 cM
Ctg-413	*AvrPm3f1*	*K413_2*	25.4 cM
Ctg-125	*AvrPm3f2*	M125RE	14.9 cM
		M125LE	10 cM
Ctg-026	*AvrPm3f2*	*K26_04*	6.2 cM
		K26_03	6.2 cM
		K26_06	6.2 cM
		K26_02	6.2 cM
Ctg-426	*AvrPm3f2*	*K426_3*	6.2 cM
		K426_4	6.2 cM
		K426_6	4.6 cM
		M426MI	4.6 cM
Ctg-052	*AvrPm3f2*	*K52_02*	1.5 cM
		K52_24	0.0 cM
		K52_09	0.0 cM
		K52_05	0.0 cM
		K52_10	0.0 cM
		K52_07	0.0 cM
Ctg-033	*AvrPm3f2*	*K33_11*	0.8 cM
		M033RE	12.3 cM

[1] BAC-contigs identified by BSA as associated with avirulence towards *Pm3f*. Genetic association was determined based on the percentage of SNP positions corresponding to the genotype of the avirulent parent 96224 (see Methods).
[2] Genetic linkage to *AvrPm3^{f1}* (*locus_1*) or *AvrPm3^{f2}* (*locus_2*) was determined based on contig-specific genetic markers indicated in the third column.
[3] Contig specific genetic markers. KASP markers are indicated by 'M' followed by contig-number and relative physical position of the marker indicated as 'LE' (Left end), 'RE' (Right end) or 'MI' (middle). CAPS markers are italicized and indicated by 'K' followed by contig-number and marker identifier.
[4] Genetic distances from *AvrPm3^{f1}* or *AvrPm3^{f2}* of contig-specific markers indicated in centiMorgans (cM).

SNP calling

Because we were unable to distinguish whether some of the SNPs in the mapping of the bulk (Figure 3A) were true or sequencing errors in either the reference or the mapped sequence, we mapped reads of both parental isolates to the reference (Figure 3B and 3C). The shared SNPs that were identical in both mappings (i.e. errors in the reference sequence) were filtered out and removed from the bulk mapping (Figure 3D). Moreover, since we were looking

for those SNPs that correspond to the allele of the avirulent parent 96224, which is also the reference sequence, we would not get a signal from those reads that match the reference for 100 % in the SNP calling, simply because they would not be considered a SNP. To circumvent this and also to receive the SNPs which we were most looking for, we included all those SNPs, which were found in the mapping of 94202 but did not show up in the bulk mapping and set the respective positions in the bulk mapping to 100% of the 96224 haplotype (Figure 3E).

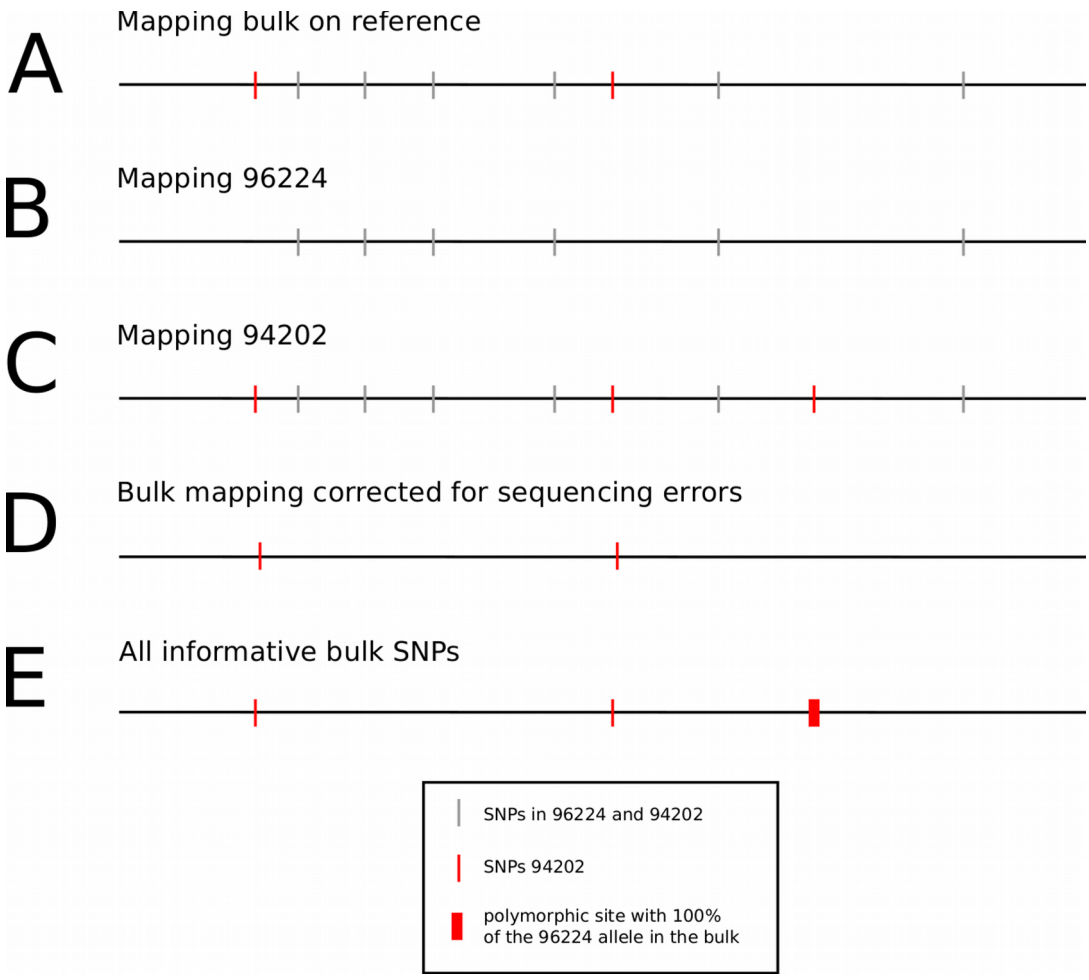


Figure 3. Schematic illustration of the criteria to identify relevant SNPs from the bulk mapping.

In total we identified 99,826 informative SNP positions (each with a certain percentage of either one of the parental genomes) in the bulk data. The histogram of the allele frequencies among all SNPs shows almost no SNPs with low frequencies of the 94202 genotype. This is considered a technical artifact of the quality filtering for TEs. However, the tail which corresponds to 100 % of the 96224 genotype and thus the avirulent phenotype is clearly enriched (Figure 4).

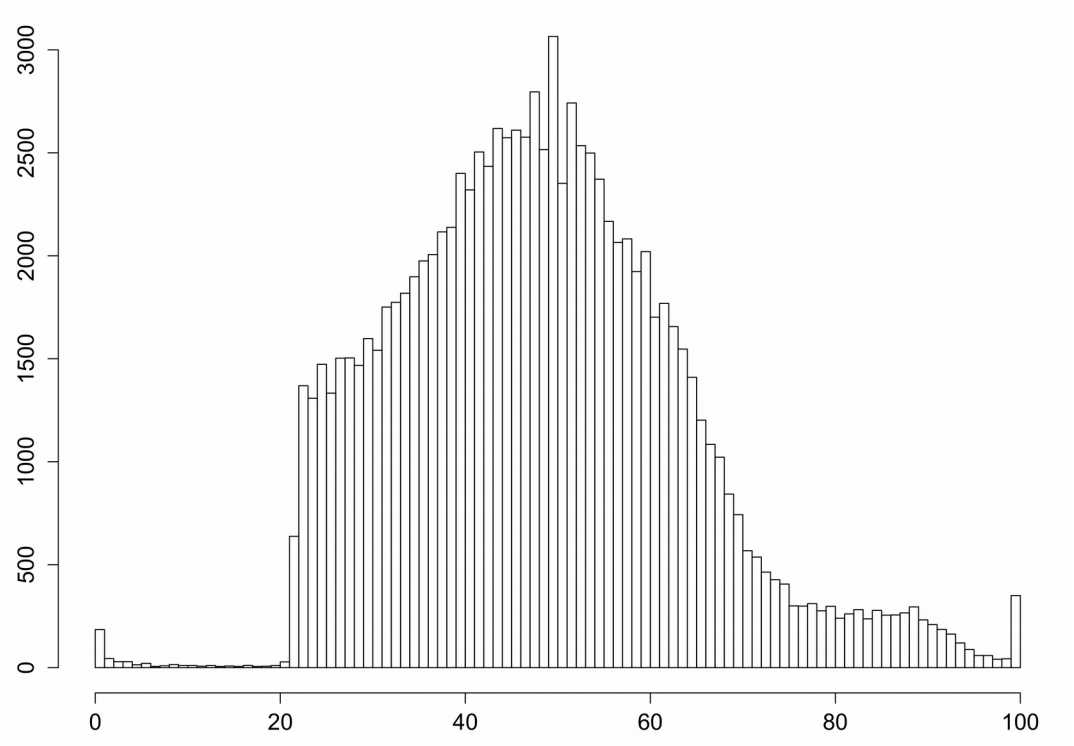


Figure 4. Distribution of the corrected SNP frequencies in the Bulk mapping. The percentage on the x-axis corresponds to the 96224 genotype. We found an enrichment of 100 % hits which are the candidate regions, corresponding to the avirulent phenotype. The sharp drop on the left side is considered a technical artifact from the previous quality filtering for TEs.

Association mapping

We further reduced the data-set by selecting only those contigs that at least contain positions with a 96224 allele frequency of > 90 %. This criterion resulted in 5413 positions on 27 contigs. As a positive control, we included Locus 1, which was associated with the 96224 haplotype and manually assembled to a continuous piece by Parlange. To explore the data, all the SNPs of the 27 contigs were plotted and visually examined. The pattern that we ideally expected to see on one of the contigs was an increasing frequency for the 96224 allele that would peak in a stretch of 100 % hits.

For Locus 1, basically all SNPs were 100 % associated with the 96224 genotype despite two noise signals (Figure 5A) which basically proofed the principle of the analysis. However, because of the relatively low number of progeny, the bulk did not provide a high enough recombination frequency to point to a single contig or even a single gene. In fact, we found three contigs that gave strong signals on their full length similar as for Locus 1 (Contigs 52, 424 and 426). Moreover, we identified several contigs that were most likely wrongly assembled in the reference assembly. In these we found stretches of strong signals that would suddenly drop to an average level of approximately 50 % (Figure 5B). Moreover, these breaks were always found associated with adjacent sequence gaps. As an example for an average region of the genome, we show a fragment of Contig 10,000, which consists of all the contigs that could not be linked with the FPC algorithm (Figure 5C). However, even though there are sporadic 100% signals, this region would not be considered as linked to the 96224 haplotype.

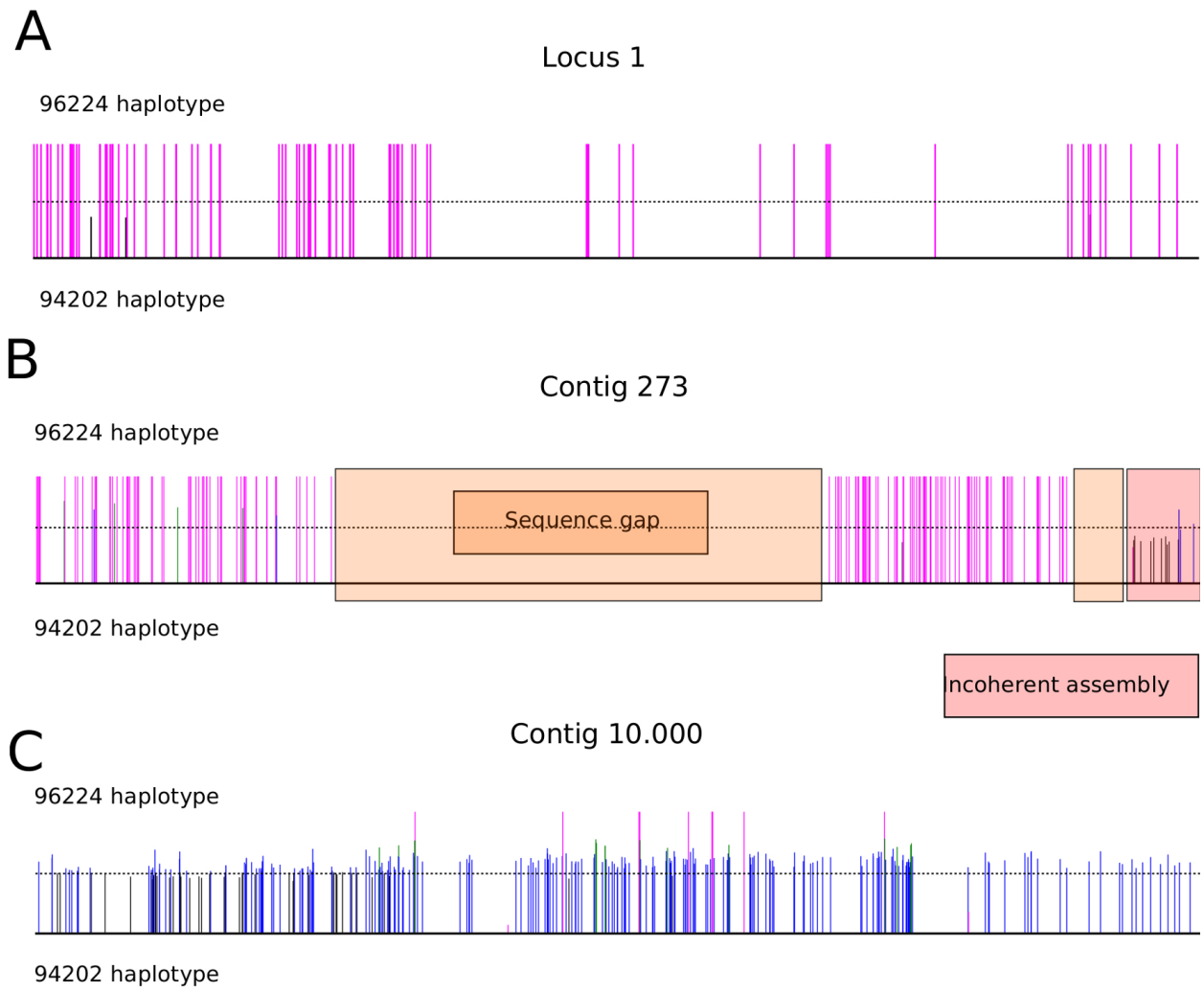


Figure 5. Plotting of the SNP ratios on to FPC contigs. A: Locus 1, identified by Parlange *et al.* (2015) which we used as positive control shows only SNPs that are 100% associated with the avirulence phenotype. B: An example for a typical FPC contig showing strong association and a putative miss-assembly indicated by the sudden drop of association (right end of the contig). C: Example for an average region with an equal distribution on both parental genotypes.

Based on these results, contig-specific cleaved amplified polymorphic sequence (CAPS) markers were designed for high-resolution mapping of Locus 2. We designed markers on nine contigs that were genetically associated to the 96224 genotype (Table 2). Additionally, one marker was designed from a continuous stretch of 100 % hits on Contig 10.000. We identified a marker on Contig 33, flanking the gene at 0.7 cM (one recombinant) on one side and at another marker on Contig 52 at 1.5 cM (two recombinants) on the other side. Moreover, we found six co-segregating markers, all from Contig 52.

BAC sequencing and annotation

To uncover the full sequence of Locus 2, we used the BAC fingerprint data from the original assembly and an algorithm that is specially designed to handle fingerprint data of repetitive genomes, the linear topological contig algorithm (LTC / Frenkel, 2010). Based on the new FPC assembly, the BACs of Contig 33 and Contig 52 could be linked. Seven BAC clones were selected to be sequenced each individually with Illumina sequencing. These sequencing reads were thereafter assembled and liked into a 445-kb contiguous sequence. 90% of this sequence turned out to be TEs but also 14 genes were identified. Among these, eight were predicted to encode for putative effector genes. These were thereafter cloned and transiently co-infiltrated with Pm3f into leaves of *Nicotiana Benthamiana*. Finally, one of these effector candidates induced cell death which demonstrating its function as an avirulence gene on Pm3f. This analysis was part of a project leading to a publication in the journal *Plant Cell* by Bourras *et al.* which was published in December 2015. Figure 6 shows a reconstruction of a pseudo-molecule of the AvrPm3F phenotype linkage group which shows a nice peak in the region where the gene was identified (Figure 6).

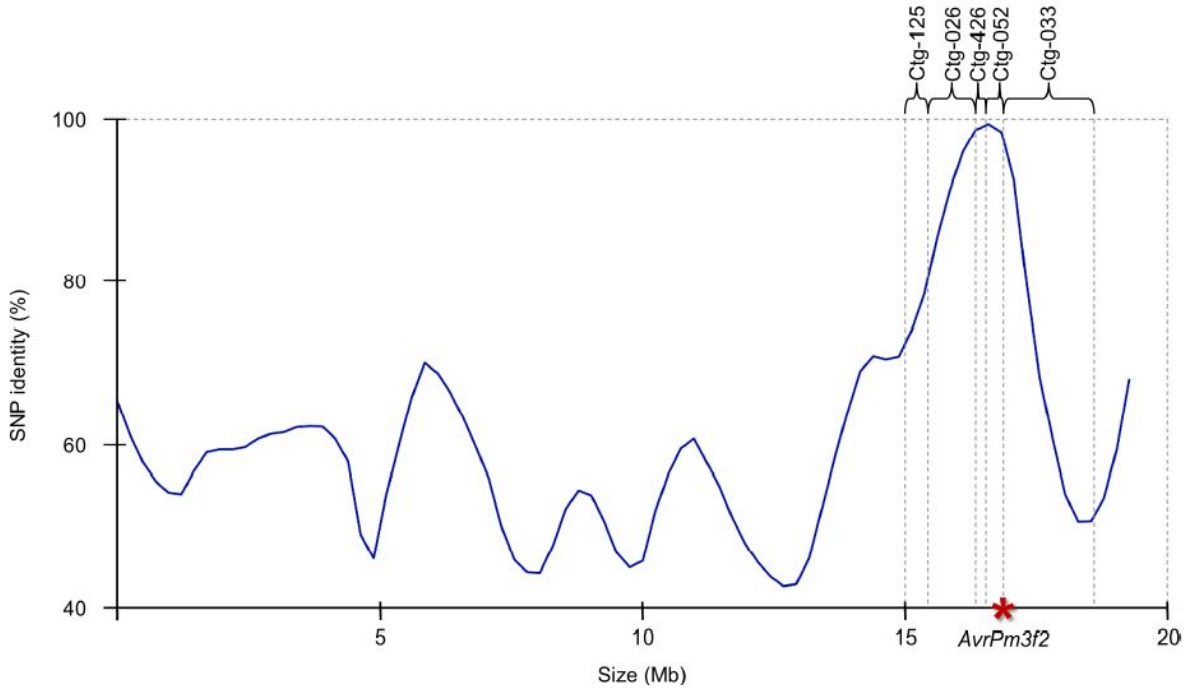


Figure 6. Pseudo-molecule of the AvrPm3F linkage group showing a peak of the 96224 genotype in the region of the avirulence gene with the respective FPC contigs (adapted from Bourras *et al.*, 2015).

2.4. Discussion

In this project we showed how the combination of different strategies can be a very powerful approach for the identification of target genes. We combined information of genetic mapping, involving different technologies such as the KASP or CAPS markers, with the concept of a BAC-library, three different sequencing technologies (Sanger, Roche/454 and Illumina) and the bulk segregant analysis, which all together lead to cloning and the functional validation the first avirulence gene in wheat powdery mildew (Bourras *et al.,* 2015). Here, I briefly discuss advantages and limitations of the bulk segregant analysis.

In principal, our approach was very successful. The identified genomic regions, were all consistent with the genetic data. However, to finally narrow down the region to one candidate gene, more sensitive strategies with a higher resolution were needed. The genetic map of the 96224 x 94202 cross comprised approximately 2,000 cM. This means the average number of crossover per progeny equals 20 recombinations per clone. Accounting for the 23 progeny mapping population that was used for the BSA, this results a total of 460 expected crossover events in the total dataset. Given the *B.g. tritici* genome size of 180 Mb divided by the number of expected recombination events of 460 equals a resolution of one recombination event every 400 kb.

As Wicker *et al.* (2013) showed, there are very conserved haplogroup blocks among the different isolates, suggesting a complex population structure. Possibly, there are recombination hot-spots that lead to the exchange of certain “blocks” or “packages” of genes. Moreover, the distribution of the allele-frequencies (Figure 5) shows abrupt drops to approximately 75 and 90 % of association, respectively, instead of an evenly flattening normal distribution. To obtain a more detailed picture on the distribution of recombination events, each of the individual progeny of the bulk could be bar-coded and re-sequenced. This would allow to identify each recombination event on the level of individuals.

Thus, only larger mapping populations and more recombination can provide the resources to map target genes to smaller sequence intervals. To generate crosses and to maintain mapping populations, however, is very time and labor

consuming. Each individual must be grown from a single spore and inoculated regularly on fresh leaf material. Also phenotyping is difficult and must be done manually. Therefore, these biological properties can be considered as the true limiting factors.

The bulk segregant analysis described here, could moreover be applied to other crosses to identify further genomic regions of interest and additional avirulence genes. At once multiple crosses and reliable phenotyping data is available, these can be incorporated into more sophisticated meta studies, Careful selection of the crossed isolates is therefore of great importance to cover a broad diversity. It will be highly interesting to study distribution of recombination along *Blumeria* chromosomes. One can speculate that effectors would preferably lie in regions with high recombination frequencies. Thus this might be another approach to identify them. Last but not least, it would be of (at least my personal) interest to also investigate the TE content of these regions to examine possible associations with recombination events.

2.5. References:

Bhullar NK, Zhang Z, Wicker T, Keller B:. **Wheat gene bank accessions as a source of new alleles of the powdery mildew resistance gene Pm3: a large scale allele mining project.** *BMC Plant Biol. 2010,* **10:**88.

Brunner S, Hurni S, Streckeisen P, Mayr G, Albrecht M, Yahiaoui N, Keller B: **Intragenic allele pyramiding combines different specificities of wheatPm3 resistance alleles.** *Plant J. 2010,* **64:**433–445.

Bourras S, McNally KE, Ben-David R, Parlange F, Roffler S, Praz CR, Oberhaensli S, Menardo F, Stirnweis D, Frenkel Z, Schaefer LK, Flückinger S, Treier G, Herren G, Korol AB, Wicker T, Keller B: **Multiple Avirulence Loci and Allele-Specific Effector Recognition Control the Pm3 Race-Specific Resistance of Wheat to Powdery Mildew.** *Plant Cell 2015*, **27(10):**2991-3012.

Denancé N, Sánchez-Vallet A, Goffner D, Molina A: **Powdery mildew fungal effector candidates share N-terminal Y/F/WxC-motif.** *Front. Plant Sci. 2013,* **4:**155.

Flor HH: **Current status of the gene-for-gene concept**. *Annu. Rev. Phytopathol. 1971,* **9**: 275–296.

Frenkel Z, Paux E, Mester D, Feuillet C, Korol A: **LTC: a novel algorithm to improve the efficiency of contig assembly for physical mapping in complex genomes.** *BMC Bioinformatics 2010*, **11:**584.

Glawe DA: **The powdery mildews: a review of the world's most familiar (yet poorly known) plant pathogens.** *Annu. Rev. Phytopathol. 2008*, **46:**27–5.

Godfrey D, Böhlenius H, Pedersen C, Zhang Z, Emmersen J, Thordal-Christensen H: **Powdery mildew fungal effector candidates share N-terminal Y/F/WxC-motif.** *BMC Genomics 2010,* **11:**317.

He C, Holme J, Anthony J: **SNP genotyping: the KASP assay.** *Methods Mol. Biol. 2014,* **1145:**75–86.

Hückelhoven R, Panstruga R: **Cell biology of the plant-powdery mildew interaction.** *Current Opinion. Plant Biol. 2011*, **14(6):**738-46.

Jones JDG, Dangl JL: **The Plant immune system.** *Nature 2006*, **444:**323-329.

Luna E, Pastor V, Robert J, Flors V, Mauch-Mani B, Ton J: **Callose Deposition: A Multifaceted Plant Defense Response.** *MPMI 2010,* **24:**183-193.

McHale L, Tan X, Koehl P, Michelmore RW: **Plant NBS-LRR proteins: adaptable guards.** *Genome Biology 2006,* **7:**212.

Parlange F, Roffler S, Menardo F, Ben-David R, McNally KE, Oberhaensli S, Stirnweis D, Buchmann G, Wicker T, Keller B: **Genetic and molecular characterization of a locus involved in avirulence of Blumeria graminis f. sp. tritici on wheat Pm3 resistance alleles.** *Fungal Genet 2015,* **82:**181-92.

Spanu PD: **The genomics of obligate (and nonobligate) biotrophs.** *Annu. Rev. Phytopathol. 2012,* **50:**91-109.

Takagi H, Abe A, Yoshida K, Kosugi S, Natsume S, Mitsuoka C, Uemura A, Utsushi H, Tamiru M,

Takuno S, Innan H, Cano LM, Kamoun S, Terauchi R: **QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations.** *Plant Journal 2013*, **74:**174–183.

Yahiaoui N, Srichumpa, P, Dudler, R, Keller B: **Genome analysis at different ploidy levels allows cloning of the powdery mildew resistance gene Pm3b from hexaploid wheat.** *Plant J. 2004,* **37:**528–538.

Yahiaoui N, Kaur N, Keller B: **Independent evolution of functional Pm3 resistance genes in wild tetraploid wheat and domesticated bread wheat.** *Plant J. 2009,* **57:**846–856.

Chapter 3:

Genome-wide comparison of Asian and African rice reveals high recent activity of DNA transposons

Here, we compared the genomes of the two rice species *O. sativa* and *O. glaberrima* to find polymorphic loci associated with the activity of DNA transposons. This work was published by Roffler and Wicker in *Mobile DNA 2015.*

Roffler and Wicker *Mobile DNA* (2015) 6:8
DOI 10.1186/s13100-015-0040-x

Mobile DNA

RESEARCH Open Access

Genome-wide comparison of Asian and African rice reveals high recent activity of DNA transposons

Stefan Roffler and Thomas Wicker*

Abstract

Background: DNA (Class II) transposons are ubiquitous in plant genomes. However, unlike for (Class I) retrotransposons, only little is known about their proliferation mechanisms, activity, and impact on genomes. Asian and African rice (*Oryza sativa* and *O. glaberrima)* diverged approximately 600,000 years ago. Their fully sequenced genomes therefore provide an excellent opportunity to study polymorphisms introduced from recent transposon activity.

Results: We manually analyzed 1,821 transposon related polymorphisms among which we identified 487 loci which clearly resulted from DNA transposon insertions and excisions. In total, we estimate about 4,000 (3.5% of all DNA transposons) to be polymorphic between the two species, indicating a high level of transposable element (TE) activity. The vast majority of the recently active elements are non-autonomous. Nevertheless, we identified multiple potentially functional autonomous elements. Furthermore, we quantified the impacts of insertions and excisions on the adjacent sequences. Transposon insertions were found to be generally precise, creating simple target site duplications. In contrast, excisions almost always go along with the deletion of flanking sequences and/or the insertion of foreign 'filler' segments. Some of the excision-triggered deletions ranged from hundreds to thousands of bp flanking the excision site. Furthermore, we found in some superfamilies unexpectedly low numbers of excisions. This suggests that some excisions might cause such large-scale rearrangements so that they cannot be detected anymore.

Conclusions: We conclude that the activity of DNA transposons (particularly the excision process) is a major evolutionary force driving the generation of genetic diversity.

Keywords: DNA transposon activity, Rice, Proliferation mechanism

Background

Transposable elements (TEs) are found in practically all eukaryotes and are thought to have co-evolved with cellular life. Due to their virus-like lifestyle, TEs are considered 'parasitic' or 'selfish' DNA. However, recent studies revealed more detail about their role as potent genome shapers [1-4]. Most generally, TEs can be divided into two major classes such as: Class I (retrotransposons) and Class II (DNA transposons). Each class is further subdivided into several superfamilies [5]. For this study, we used the proposed classification system where each superfamily was assigned a 3-letter code [5] which will be given in parenthesis.

Retrotransposons use a mRNA intermediate that is reverse transcribed and integrated somewhere else in the genome. Therefore, each successful transposition produces an additional copy, which can lead to massive genome expansions [2,6]. In contrast, DNA transposons use a *cut and paste* mechanism to transpose and multiply. Because DNA transposons of the terminal inverted repeat (TIR) order [5] are the main focus of this study, we will describe their characteristics in more detail. In most TIR superfamilies, the pivotal transposase is flanked by TIRs and is transcribed and translated by the host machinery. *Mariner* (*DTT*) elements are, in copy numbers, the most abundant DNA transposons in rice and other grasses such as Sorghum [7] or Brachypodium [8] and usually encode a single transposase protein containing a catalytic DDD/E motif as do elements of the *hAT (DTA)* and *Mutator (DTM)* superfamilies. In contrast, *Harbinger*

* Correspondence: wicker@botinst.uzh.ch
Institute for Plant Biology, University of Zürich, Zollikerstrasse 107, CH-8008 Zürich, Switzerland

(DTH) and *CACTA (DTC)* elements also encode a second open-reading frame (ORF) of yet unknown function. Additionally, *CACTA* elements often contain complex arrays of subterminal repeats and large arrays of low complexity repeats which make them difficult to assemble and annotate [9].

Many transposons have lost their ability to transpose on their own. These non-autonomous elements usually lack protein-coding domains and transpose by recruiting enzymes of active, full size 'mother' elements. Such trans-acting systems have been described in both DNA [10] and retrotransposons [11]. Often, the non-autonomous elements, by far, outnumber their full-length counterparts and can represent a substantial amount of DNA in some genomes [8,12]. For example, in *Brachypodium*, 20,994 non-autonomous *Mariner* (*DTT*) elements were found whereas only 50 putative mother elements were identified [3]. Small non-autonomous DNA transposons are often referred to as 'Miniature Inverted Transposable Elements' (MITEs, [13,14]). Since we found non-autonomous elements of various sizes in multiple superfamilies, we prefer not to use the term MITE but rather refer to them simply as non-autonomous elements.

So far, several active DNA transposons have been found and documented in rice. The first described element was the non-autonomous, low copy element *mPing* of the *Harbinger (DTH)* superfamily [15-17]. One study identified *mPing* through mutability of a slender mutation of the glume which was caused by the insertion of *mPing* into the *slg* locus [15]. Kikuchi *et al*. identified *mPing* by a computational approach and presented a putative corresponding autonomous element which they named *Ping* [16]. Moreover, they showed experimentally that the transposition of both *mPing* and *Ping* preferentially occurs in cells derived from germ-line cells. Jiang *et al*. identified an additional, more distantly related autonomous element (*Pong*) which can activate *mPing in trans* [17]. Moreover, they could show experimentally that *mPing* preferably inserts in single-copy sequences. The *mPing*/*Pong* system has later been shown to transpose when introduced in heterologous systems such as in yeast [18] or Arabidopsis [19]. In 2005, Fujino *et al*. identified a non-autonomous element of the *hAT (DTA*) superfamily, *nDart*, that causes an *albino* phenotype and its putative autonomous mother element, *Dart*, which shared identical TIRs and similar subterminal sequences [20]. Finally, another member of the *hAT (DTA)* superfamily, *dTok*, was found to have inserted into the kinase domain of *FON1* during the molecular analysis of the *fon1/mp2* mutant [21]. Also here, they propose a putative autonomous element providing the necessary enzymes for the mobility of *dTok*. Interestingly, also in this study, transposon activity was found only in regenerative tissue.

Upon insertion, the host's DNA is cut similar to a restriction enzyme, generating 3′ overhangs. After the transposable element (TE) has been inserted, these overhangs get complemented by the host's repair system on both sides of the TE which leads to a duplication of the original target site. The length of this target site duplication (TSD) is an important diagnostic feature to classify DNA transposons, especially non-autonomous ones which do not encode any proteins (Table 1).

The current model of transposon excision proposes initial binding of the transposase to the TIR sequences followed by sequential cleavage of the two DNA strands. Thereon, dimerization of the paired-end complex brings the two strands in close proximity and links them by a clamp-loop protein [22]. Most likely, at least two subunits of the transposase (one binding to each TIR) are required for cleavage at the border of the element. When DNA transposons excise, they leave a double-strand break (DSB) with small 3′ overhangs which are derived from the TIRs of the element [22] (Additional file 1: Figure S1). Since DSBs are lethal for dividing cells, they need to be repaired by the host's DSB repair systems. The applied repair pathways and therefore the footprint of the excision can vary substantially between species. There are two main groups of DNA repair pathways [23-25]. Which of the different pathways is applied depends on the cell-cycle phase and the nature of breakpoint ends. The simplest way of DSB repair is that the 3′ overhangs get denatured by exonucleases. This generates blunt ends which allow direct ligation of the two strands, called non-homologous end joining (NHEJ). These cases result in what is referred to as 'perfect excision' where only the TSD remains as a footprint [3] (Additional file 1: Figure S1A). The second major pathway uses short homologous sequences as templates to connect the two strands. These processes employ exonucleases to produce 5′ overhangs which resect until the newly exposed strands find a homologous region of a few bp between each other, allowing annealing of the overhangs. This is referred to as microhomology-mediated end joining (MMEJ) or single strand annealing (SSA). As a consequence, the sequence downstream of the homology will be lost resulting in a deletion (Additional file 1: Figure S1B). In some cases, if the homologous pattern that re-ligates the two

Table 1 Target site characteristics of DNA transposon superfamilies

TE superfamily	Target site motif	Target site size	TIR consensus
Mariner (DTT)	TA	2	CTCCCTC
Harbinger DTH)	TAA/TTA	3	GG(G/C)CC
Mutator (DTM)	Variable	9	GAG
CACTA (DTC)	Variable	3	CACT(A/G)
hAT (DTA)	Variable	8	CA

strands corresponds exactly the complementary target site, this can lead to a restoration of the initial, 'empty-site' situation even before insertion of the TE (Additional file 1: Figure S1D). Such 'precise' excisions have been described to occur frequently when introducing the *mPing*/*Pong* system into Arabidopsis [19]. Thus, it is important to note that precise excisions are indistinguishable from insertions purely by means of comparative analysis. Alternatively, ectopic recombination can be initiated, which is referred to as synthesis-dependent strand annealing (*SDSA*). This can lead to the introduction of copies of foreign segments as 'filler' DNA (Additional file 1: Figure S1C). SDSA is also the mechanism underlying gene conversions [26]. In some cases, combinations of SSA and SDSA are utilized at the breakpoint leading to chimeric repair patterns [25]. While TE insertions or precise excisions are relatively easy to identify (*via* TSD), in some cases, it can be very difficult to precisely decipher excision footprints. Buchmann *et al.* [3] suggested that excisions of DNA transposons often cause extensive deletions which may also be combined with the introduction of foreign filler DNA.

In this work, we compared the genome sequences of Asian rice, *Oryza sativa ssp. japonica*, and African rice, *Oryza glaberrima*, whose genome sequence recently became available [27]. They diverged only about 600,000 years ago, providing an excellent opportunity to study recent TE activity and fixation. Moreover, it provided insight into the insertion and excision footprints, allowing inferring of qualitative and quantitative differences between TE superfamilies populating the two rice genomes. We aligned more than 63% of the two genomes and investigated 1,821 polymorphisms manually. Among these, we identified 487 loci with polymorphic DNA transposons that either inserted or excised since the divergence of the two species. We therefore estimate that the two rice genomes contain approximately 4,000 such polymorphisms. Moreover, we found differences in the excisions between different TE superfamilies. These seem to cause a multitude of rearrangements; some may be so dramatic that they cannot be detected at all anymore.

Results

TE families are unequally distributed within superfamilies

The assembled genome sizes (excluding Ns) are 372 Mbp for *O. sativa* (Version 5) [28] and 303 Mbp for *O. glaberrima* [27]. We were able to align approximately 63% of the two genomes (see below). We focused this study on DNA transposons of the TIR order (that is, elements are flanked by terminal inverted repeats and move with the help of a transposase enzyme). To obtain an overview of the abundance of DNA transposon families that had been active since the divergence of *O. sativa* and *O. glaberrima*, we used a database that was created based on an iterative search of insertion/excision polymorphic sequences in the alignments of the two genomes (see below and 'Methods'). Thus, the results of our survey do not reflect the total content of DNA transposons in rice which, in fact, might be much higher [27]. We identified 64,645 Class II transposons of the TIR order in *O. sativa* and 54,280 in *O. glaberrima*, occupying approximately 20.4 Mbp and 12.6 Mbp of the two genomes, respectively (Additional file 2: Table S1). The average sizes of 316 bp and 230 bp reflect the strong outnumbering of autonomous by non-autonomous elements. A closer investigation of the substantial number of unclear sites (Ns) indicated that many sequence gaps in the *O. glaberrima* assembly are caused by Class II transposons (see 'Methods'). We estimate that at least 5,100 sequence gaps actually correspond to Class II TIR elements, resulting in an estimated total of approximately 59,500 elements (approximately 16 Mbp) in *O. glaberrima*. Therefore, the overall DNA transposon content in *O. glaberrima* is probably slightly lower than in *O. sativa.*

In both species, the highest copy numbers were found for *Mariner* (*DTT*) elements, followed by elements of the *Harbinger (DTH)* superfamily. CACTA elements are on average larger than the other superfamilies (938 bp in *O. sativa* and 600 bp in *O. glaberrima*) and thus they occupy the most space. These findings are also consistent on the family level (that is, among elements that can be aligned at the DNA level [5]). We found strong over representation of a few families that dominate each of the superfamilies (Figure 1 and Additional file 2: Table S1). With the exception of the *DTA_Coraline* family (which was only found in *O. sativa*), all TE families are represented at similar numbers in both genomes (Additional file 2: Table S1).

Identification of transposon polymorphisms

We were able to align 63.3% (235.6 Mbp) of the *O. sativa* and *O. glaberrima* genomes with the Smith-Waterman algorithm in sliding windows of 12 kb (see 'Methods') and examined the presence/absence of polymorphisms larger than 50 bp. We identified 23,709 polymorphisms in the *O. sativa* genome of which 7,542 showed homology to DNA transposons. In *O. glaberrima*, we found 22,003 polymorphisms whereof 4,816 had homology to DNA transposons. Upon visual inspection, we noticed that many of the polymorphisms in *O. glaberrima* that showed homology to TEs also contained large stretches of Ns, indicating that TEs are often problematic to assemble completely. Thus, in an independent approach, we estimated how many of the 'presence' polymorphisms in *O. glaberrima* which are comprised mostly of Ns actually correspond to TE sequences (see 'Methods'). We estimate that there are approximately 1,750 polymorphisms in *O. glaberrima* which can be attributed to DNA transposons but which are not identifiable because the sequence assembly is incomplete

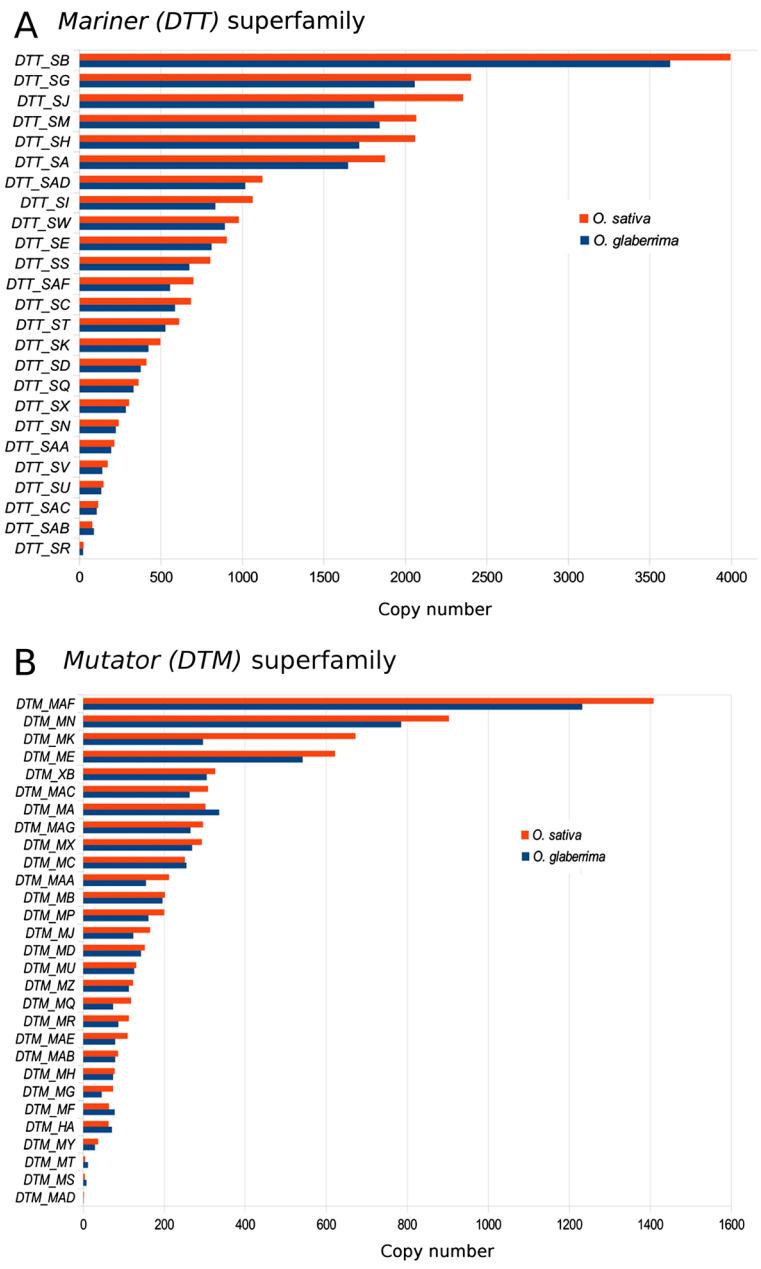


Figure 1 The abundance of *Mariner* (*DTT*) and *Mutator (DTM)* families in *O. sativa* and *O. glaberrima*. **(A)** Overview of *Mariner (DTT)* abundance. Copy numbers of individual families show large differences within the *Mariner (DTT)* superfamily. For example, in *O. sativa*, the most successful *DTT_SB* is represented 4'702 times while we only identified 25 copies of the *DTT_SR* family. **(B)** Overview of *Mutator (DTM)* superfamily. Despite an overall similar distribution, we found one exception for the *Mutator (DTM)* superfamily *DTM_MA*, where we found slightly more elements in the *O. glaberrima* genome (302 copies in *O. sativa* and 336 copies in *O. glaberrima*).

at these sites. Thus we extrapolate that there are approximately 6,500 presence polymorphisms in *O. glaberrima*, slightly fewer than the 7,542 presence polymorphisms in *O. sativa* (see 'Methods').

Here, it should also be noted that sequence alignments of large genomic regions often contain misalignments caused for example by the presence on non-homologous segments or by sequence gaps in one of the species. Automated examination of sequence alignments can therefore yield very noisy data. Thus, we decided to manually analyze a subset of the identified polymorphisms. In total, we manually analyzed 1,821 cases which showed homology to TEs, 844 from *O. sativa* and 977 from *O. glaberrima*, representing approximately 15% of all TE-related polymorphisms. Most of them turned out not to be directly associated with TE activity because many

represent internal or partial deletions within the elements, which means that the missing sequence obviously was not caused by an insertion or excision of the respective transposon but by a mechanism unrelated to its activity (for example, template slippage, Additional file 3: Figure S2).

In *O. sativa*, we found 238 and in *O. glaberrima* 249 TE polymorphisms that were most likely caused by DNA transposon activity. A complete overview of all active transposons is provided in Additional file 4: Table S2. Thus, 28% and 25% of all presence/absence polymorphisms examined represent likely transposition events. For *O. sativa*, we therefore extrapolate that about 2,100 of the TE-related polymorphisms are actually caused by TE activity. The value for *O. glaberrima* is probably similar. Considering the estimated, unknown part of approximately 450 additional TE-related polymorphisms, we expect a slightly lower overall activity of 1,650 transposition events in *O. glaberrima*.

In *O. sativa*, the most abundant were *Harbinger (DTH)* and *Mariner (DTT)* elements with 95 and 90 transpositions, respectively. Moreover, we identified 33 *Mutator (DTM)* and 15 *CACTA (DTC)* elements to have transposed recently. Finally, we found five elements of the *hAT (DTA)* superfamily. Also in *O. glaberrima*, *Harbinger* (*DTH*) and *Mariner* (*DTT*) were the most prominent superfamilies with 110 and 102 transpositions, respectively. Additionally, we identified 32 *Mutator* (*DTM*), four *hAT* (*DTA*), and a single *CACTA* (*DTC*) transposition (Table 2).

Distinguishing insertions and excisions

We defined TE insertions in the classic way as follows: one species contains the TE flanked by the two direct repeats created upon insertion (the TSD) while in the other species, the TE is absent and only one copy of the TSD is present (example in Figure 2A). Of the total 487 TE polymorphisms we identified in *O. sativa* and *O. glaberrima*, we classified 393 as insertions (192 in *O. sativa* and 201 in *O. glaberrima*). It is important to note that a precise excision (that is, one that removes the TE plus one target site) cannot be distinguished from an insertion with these criteria.

Table 2 Overview of recently active DNA transposons in O. sativa **and** O. glaberrima

TE superfamily	*O. sativa*	*O. glaberrima*
Mariner (DTT)	90	102
Harbinger (DTH)	95	110
Mutator (DTM)	33	32
CACTA (DTC)	15	1
hAT (DTA)	5	4

Excisions are much more complex to identify and show various patterns of DSB repair. The simplest case, a perfect excision, was defined as an event where the TE excises and exactly leaves the two copies of the TSD as a footprint [3]. Of a total of 94 putative excision events, we identified only eight perfect excisions (example in Figure 2B). In all other cases, the excision went along with deletions of flanking sequences or the insertion of filler sequences, or both. In 43 excisions, we found that sequences flanking the element were deleted. Excluding one extreme case (see below), on average approximately 18 bp of flanking sequences were deleted per excision event (example in Figure 2C). On the other hand, in 58 cases, excision also went along with the introduction of foreign DNA segments. On average these fillers had a size of 13 bp, ranging in size from 1 to 123 bp (example in Figure 2D). Nine cases showed both deletions and introduced filler segments. The cumulative length of all deleted sequences is 926 bp while the combined length of all filler segments is 880 bp.

The most extreme case was a putative excision of a *Mariner* element of the *DTT_SC* family. Its excision went along with the deletion of a 2,479 bp fragment on one side of the element (Figure 3). We are confident that this deletion was indeed the result of the excision because the left border of the excised fragment coincides precisely with the left end of the *DTT_SC* element (Figure 3). It is highly unlikely (however, not impossible) that a random deletion would have its one breakpoint exactly at the terminus of the TE. If this case is included in the overall calculation, a total of 3,405 bp were deleted in the 94 excision events.

The ratio of insertions and excisions indicates differences in transposition between superfamilies

Inferring recent activity of DNA transposons from the numbers of excisions and insertions is not trivial. Intuitively, one would assume that the ratio of insertions and excisions is 1:1, because each excising element would simply insert somewhere else in the genome. The current hypothesis is that DNA transposons can excise during DNA replication and transpose in front of the replication fork to create an additional copy. This results in two different gametes, one with one copy and one with two copies. If the number of transposition events is large, overall equal numbers of loci derived from the two gamete types should be passed to offspring. Thus, if all the observed insertion/excision ratios in a cross-species comparison such as the one presented is considered, the ratio is actually expected to be 2:1 (Additional file 5: Figure S3). However, it is also possible that transposition happens at other points during the cell cycle, which would not lead to a replication of the respective TE. Considering this, one would therefore expect a ratio somewhere between 1:1 and 2:1 but

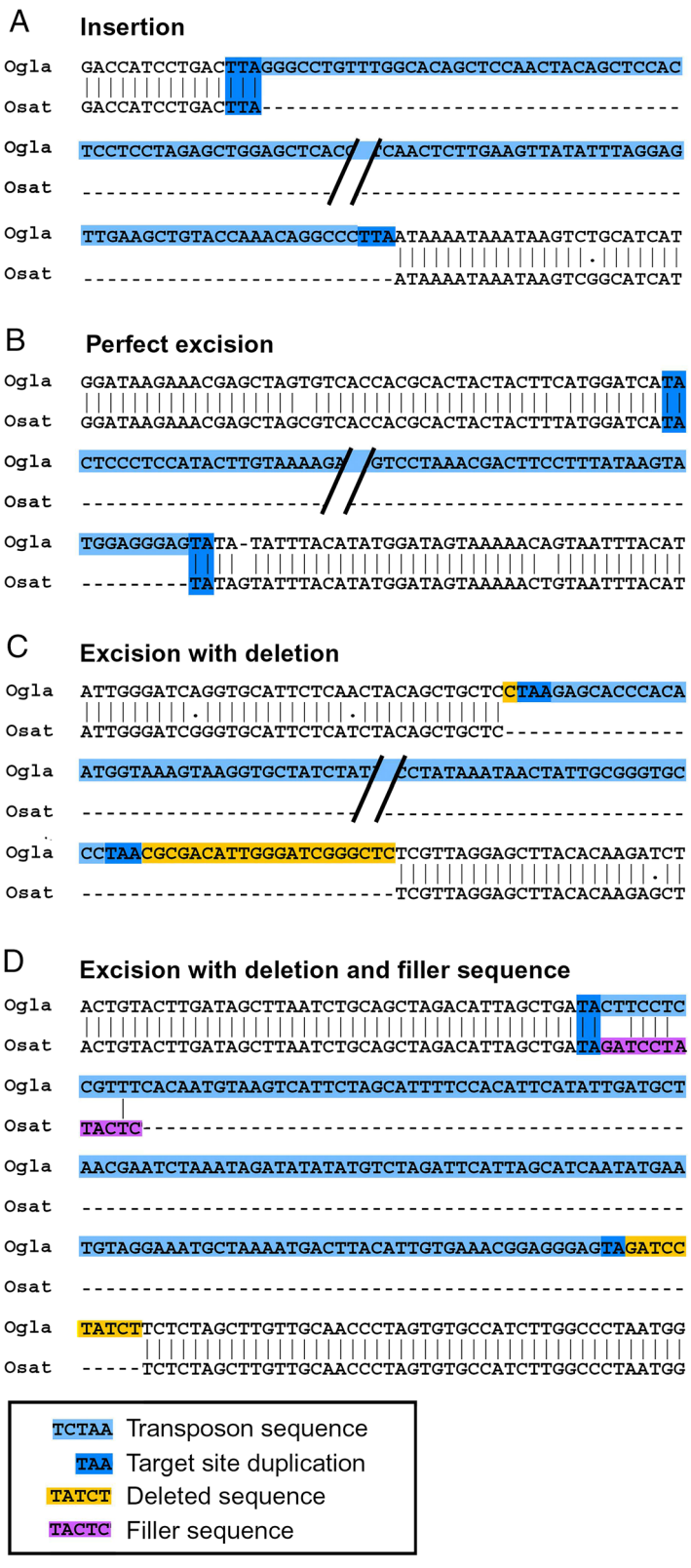


Figure 2 Examples of DNA transposon polymorphisms in *O. sativa (Osat)* and *O. glaberrima (Ogla)*. The alignments show the polymorphic TE plus some of the genomic flanking sequences. Diagnostic sequence motifs are *highlighted* with colors. **(A)** Insertion. **(B)** Perfect excision. **(C)** Excision with deletion. **(D)** Excision with deletion and filler sequence.

Figure 3 An example of a transposon excision that caused a large deletion in its flanking region. The transposon *DTT_SC* is indicated by a *gray box*. *Solid lines* represent the genomic sequences of *O. sativa* and *O. glaberrima*. The excision is precise at the left border of the *DTT_SC* element while a 2,479-bp segment was deleted at its right border.

not higher than 2:1. This 2:1 ratio is only expected if a given transposon family is active at similar levels in both species being compared. Deviation from the 2:1 ratio in the inter-species comparison would therefore indicate different levels of activity of that transposon family in the two species (Additional file 6: Figure S4). For example, a ratio much lower than 2:1 in *O. sativa* and much higher than 2:1 in *O. glaberrima* could indicate that a given transposon family was more active in *O. glaberrima* (Additional file 6: Figure S4).

When comparing the insertion/excision ratio for the different superfamilies, we observed almost the expected 2:1 ratio for the *Mariner (DTT)* superfamily. In both datasets, we found ratios which are not significantly different from 2:1 (2.6:1 in the *O. glaberrima* dataset and 2.8:1 in the *O. sativa* dataset), indicating that the proposed proliferation mechanism is sufficient to explain the observations on *Mariner* elements. It also indicates that precise excisions (removal of the TE plus one target site) are rare in the *Mariner* superfamily. Interestingly, for the *Harbinger* (*DTH*) superfamily, the ratio differs from what we expected. The ratio in the *O. glaberrima* dataset was 8.2:1 (Fisher's exact test, $P = 0.0001$), and in the *O. sativa* dataset, we found 4.4:1 ($P = 0.013$). These differing ratios between *O. glaberrima* and *O. sativa* could indicate a higher level of *Harbinger* activity in *O. glaberrima*. Finally, we found a significant difference for the *Mutator* superfamily in the *O. glaberrima* dataset where we observed a ratio of 8.7:1 with 26 insertions and only three excisions ($P = 0.03$). The ratio in the *O. sativa* dataset (4.3:1), where we found 30 insertions and seven excisions, did not reach significance level ($P = 0.14$) (Table 3). It is not easy to explain why *Harbingers (DTH)* and *Mutator (DTM)* elements deviate so strongly from the expected 2:1 ratio in both species (see 'Discussion').

Table 3 Overview of TE insertions and excisions by superfamily and species

TE superfamily	Species	Insertions	Excisions	Ratio	*P* value
DTT	*O. sativa*	59	21	2.81	0.19
DTH	*O. sativa*	84	19	4.42	0.01*
DTM	*O. sativa*	30	7	4.29	0.14
DTC	*O. sativa*	14	1	14	-
DTA	*O. sativa*	5	0	-	-
DTT	*O. glaberrima*	81	31	2.61	0.38
DTH	*O. glaberrima*	90	11	8.18	0.00008*
DTM	*O. glaberrima*	26	3	8.67	0.028*
DTC	*O. glaberrima*	0	1	-	-
DTA	*O. glaberrima*	4	0	-	-

*Significantly different from expected 2:1 ratio.

High abundance does not necessarily correlate with strong activity

To estimate activities of individual TE families, we had to consider that an additional sequence in *O. sativa* could mean that the TE inserted in *O. sativa* or excised in *O. glaberrima* and *vice versa*. Therefore, we had to combine data from both datasets. We defined the relative activity as the number of copies that moved in relation to the total copies in a particular TE family. As relative abundance we defined the total copy number of the respective family divided by the total number of DNA transposons of the investigated genome. We divided the families into three categories (Figure 4). In the first group, we grouped TE families with high overall copy numbers and also high numbers of insertion polymorphisms. Members of the *Mariner (DTT)* and *Harbinger (DTH)* superfamilies are most prominent in this category. Moreover, the most abundant *CACTA* family, *DTC_Calvin* (4,868 copies), turned out to be also very active with nine identified insertions and one excision.

The families in the second group show a high number of insertions relative to their abundance. Most noticeably, for the *DTH_TR* family, of which we found only 581 copies in the whole *O. sativa* genome, we identified 17 insertions and one excision, more than for any other family overall. Other highly active families in this class are the *Harbinger* family *DTH_TAA* and the *Mutator* family *DTM_MA* which both inserted eight times and excised once while we found 288 and 302 copies, respectively. Furthermore, we found several other *Harbinger (DTH)* and two *hAT (DTA)* families with less than 50 copies and one insertion. The most extreme case here is the *Mutator* family *DTM_MAD* where we only found two copies in the whole genome, one of them inserted recently.

The third group contains families with high abundance but with only little or even no activity. Here, we find the most numerous families, again of the *Harbinger (DTH)* and *Mariner (DTT)* superfamilies, where we found several families with more than 3,000 copies but only five or less polymorphisms. For the most numerous, *Mariner* family *DTT_SB* (3,995 copies), and the most abundant, *Mutator* family *DTM_MAF* (1,408 copies), we did not

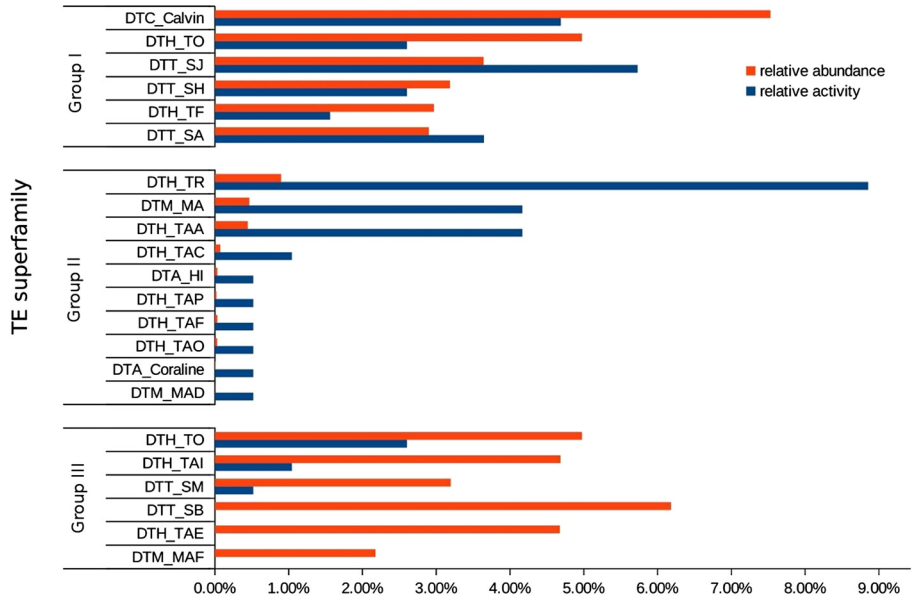


Figure 4 The relative activity and relative abundance of the TE families in *O. sativa*. We compared the relative activity with the relative abundance of all TE families in *O. sativa*. Group I consists of families with high activity and high abundance. The *CACTA* family *DTC_Calvin*, which is the overall most abundant family, also shows remarkable activity. Group II contains elements with high activity but low copy numbers. We found that *Mutator (DTM)* and *hAT (DTA)* families are relatively active despite their poor abundance. Finally, Group III consists of families with high abundance but relatively little activity. This class is dominated by families of the *Harbinger (DTH)* and *Mariner (DTT)* superfamilies. The *Harbinger* family *DTH_TO* seems to be still relatively active despite its high abundance, whereas the most abundant *Mariner* and *Mutator* families *DTT_SB* and *DTM_MAF*, respectively, show no activity at all.

find any polymorphic elements at all (Additional file 2: Table S1 and Additional file 4: Table S2).

Most potentially active and autonomous elements are of the *Mutator* superfamily

As mentioned above, the majority the elements that transposed since the divergence of the two rice species are non-autonomous elements which do not code for any proteins. We found a total of 17 elements which contain at least parts of transposase ORFs and have moved since species divergence. Interestingly, twelve of them belong to the *Mutator (DTM)* superfamily, which had, overall, relatively few active elements (see above). We found eight families where all polymorphic copies contain at least parts of coding sequences (CDS) for transposases (*DTM_MAF*, *DTM_MS*, *DTM_MAG*, *DTA_Coraline*, *DTA_HL*, *DTH_TAG*, *DTH_TAH*, and *DTH_Blip*). For the *Mutator* family *DTM_MU*, we found one CDS-containing element and one non-autonomous deletion derivative. However, besides the *DTA_Coraline* insertion, all the transposase ORFs of the above families have either stop codons or frameshifts, suggesting that they are not functional autonomous elements.

The most interesting *Mutator* family is *DTM_MK* which contains 14 elements that have moved since species divergence (Figure 5). We found a total of nine insertions and one excision of *DTM_MK* elements in *O. sativa* and four excisions but no insertion in *O. glaberrima*, indicating that they had been active in both species. Here, we found three elements with apparently intact transposase ORFs which we consider potentially active mother elements. The largest among these (6,721 bp) contains an intact transposase ORF and an additional ORF that encodes a 'TE-associated' protein. Interestingly, we found large parts of the same second ORF in three other putative full-length elements that all have disrupted transposase ORFs. Furthermore, two of the disrupted elements acquired an additional sequence which has no homology in any of the other family members (Figure 5). Intriguingly, we also found a subpopulation of six non-autonomous elements that had moved. These elements are very similar to each other in size (604 bp to 684 bp) and share the TIRs of around 120 bp with the other larger elements. The approximate 400 bps between their TIRs is not homologous to any of the larger elements but is highly similar in the six small copies. This indicates that these six copies originated from a single deletion event and multiplied after (Figure 5).

TE activity mainly influences regions close to genes but not coding sequence

We investigated if and to what extend TE activity affects genes by using the coding sequence provided by the Rice Genome Annotation Project [29]. We included all TE

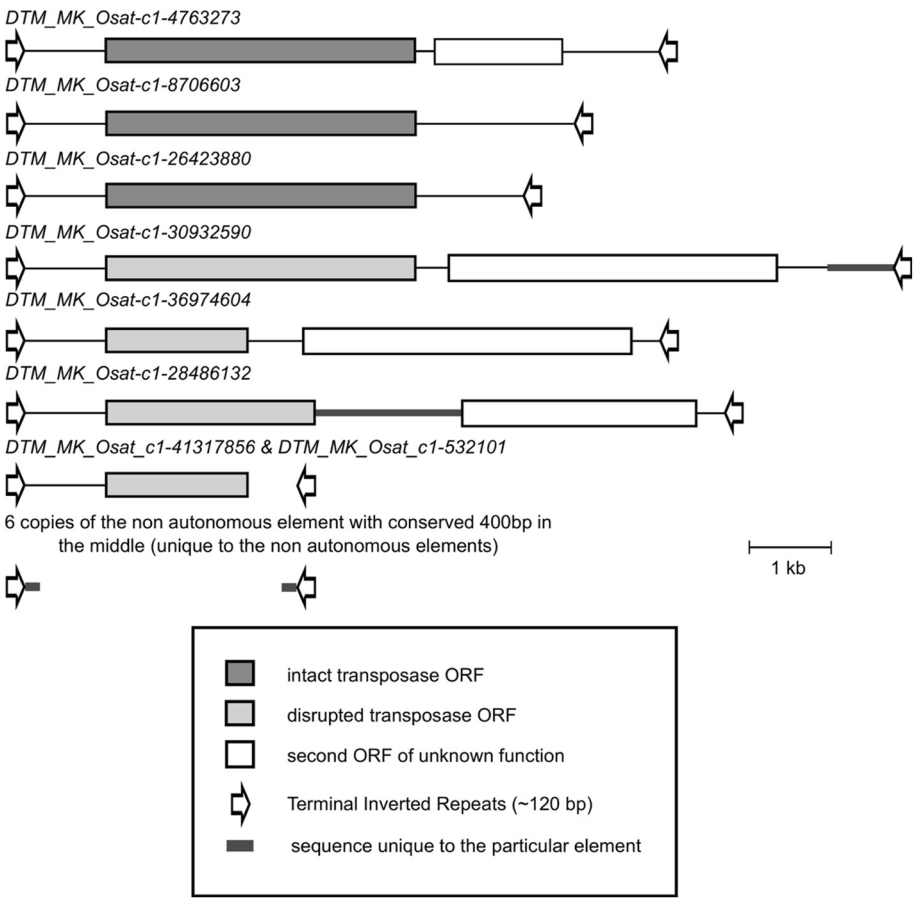


Figure 5 The schematic representation of the copies of the *Mutator* family *DTM_MK* which were polymorphic in *O. sativa* and *O. glaberrima*. The family includes three copies which contain intact transposase ORFs (top three copies). One of these putative mother elements additionally carries a fragment of a second ORF which was also found in other derivatives. Presumed non-autonomous copies have partially deleted or disrupted reading frames containing stop codons or frameshifts in the transposase ORF. Additionally, we found six copies of non-autonomous elements which consist only of TIRs plus an internal sequence that has no homology to that of larger elements (bottom). The fact that all six are very similar to each other indicates that they originate from the same deletion event and are multiplied later.

polymorphisms (insertions and excisions) found in exons as well as those in introns, 1,000 bp upstream and 500 bp downstream of the coding sequence.

Of the 487 investigated TE polymorphisms, 160 matched our criteria. We found 74 insertions or excisions in upstream regions and 24 downstream of genes. Moreover, 61 polymorphisms were identified in introns. Interestingly, only one, a *Mariner* element of the *DTT_SG* family, actually disrupted a gene in *O. sativa*. The element inserted into the second exon of a glutathione S-transferase homolog (LOC_Os01g72120), a protein assumed to be involved in detoxification. We furthermore performed a gene ontology analysis. This revealed that genes involved in nucleoside metabolic and biosynthetic processes, protein dephosphorylation and SRP-dependent proteins targeting the membrane, are affected disproportionately high ($P < 0.01$/not shown).

Discussion

In this study, we conducted a genome-wide analysis of the activity of DNA transposons in the two closely related rice species *O. sativa* and *O. glaberrima*. Numerous studies have described the activity of (Class I) retrotransposons in plants [2,6,30,31], but only very few have focused on DNA (Class II) transposons. To our knowledge, this is the first study that characterizes DNA transposons and assesses their activity at a genome-wide scale. Here, the recently released sequence of the *O. glaberrima* genome [27] provided a unique opportunity for comparative analysis because it is phylogenetically close enough to *O. sativa* to allow reliable sequence alignments of large parts of the genome and yet distant enough to have accumulated numerous TE polymorphisms. Both the *O. sativa* and the *O. glaberrima* genome were sequenced with Sanger technology and assembled independently. This has the important

advantage over simple re-sequencing and subsequent mapping onto a reference so that large insertions and deletions in both species can be easily identified and characterized in much detail. A very important part of our work was that we manually inspected over 1,800 TE related polymorphisms in the two species because transposon excisions, especially, can produce complex sequence patterns which are extremely difficult to characterize on an automated basis. Furthermore, it was important to distinguish actual insertions and excisions from random deletions that by chance affected parts of the TEs. The result of this study is a chromosome-scale catalog of TEs that were recently active in rice as well as information on how transposon insertions and excisions affect the genome. Our data allow conclusions and hypotheses on transposon activity. These will be discussed below.

Estimates of frequencies of transposition events in *O. sativa* and *O. glaberrima*

It has been known for several years now that grass genomes contain tens of thousands of DNA transposons [7,12]. However, it was not clear how often these elements actually move. *Wessler et al.* [10] suggested that some families may be active in bursts, creating thousands of copies within only a few generations before they are silenced by the host. The data from our study now allow some conclusions on the actual level of transposition activity such that:. *O. sativa* and *O. glaberrima* were estimated to have diverged approximately 600,000 years ago [27]. Overall, our data indicate that DNA transposons were active at similar levels in both rice species since their divergence. Based on the manual analysis of insertions, we estimated that *O. sativa* contains roughly 2,000 polymorphic elements and *O. glaberrima* approximately 1,600. Assuming that most of these polymorphic transposons are actually fixed in the two species, we can estimate that since species divergence, a DNA transposon polymorphism (insertion or excision) became fixed approximately every 250 years to 300 years in both *O. sativa* and *O. glaberrima*. In other words, about 2.5% to 3.5% of the DNA transposons in the two species have moved within the last 600,000 years.

For the following calculations, we assume that all identified transposition events were selectively neutral as deleterious transpositions would have been selected against. However, fixed polymorphisms only represent a small part of actual TE activity. A measure for actual transposon activity can be defined analogous to a mutation rate (m) as the number of transposition per generation per individual. The total number of transposition events per generation would therefore be the effective population size N(e) times the mutation rate (that is, N(e)m). According to Kimura [32], fixation rates are inversely proportional to population sizes. Thus, if all transposition events are neutral, the probability of fixation of an event is 1/N(e). The rate of fixation is therefore N(e)m × 1/N(e) = m. Thus, population size is irrelevant, and the fixation rate is equal to m [32]. In the case of *O. sativa*, fixation rate would therefore be the number of identified transposition events (assuming all of them are fixed) divided by the number of generations since divergence from *O. glaberrima* (2,300/600,000 = 0.004). This would mean that in each generation, 1 out of 250 individuals contains a transposition event.

Most polymorphic DNA transposons are non-autonomous, except in two superfamilies

The vast majority of the polymorphic transposons were small non-autonomous elements (MITEs) of the *Mariner (DTT)* and *Harbinger (DTH)* superfamilies. Interestingly, we did not find any polymorphic potentially autonomous elements for either of the two superfamilies. This could indicate that the required transposase genes may still be expressed, but the mother elements themselves have lost the ability to move. It was previously reported that non-autonomous *Mariner* and *Harbinger* elements could also be cross-activated by even distantly related mother elements and even in heterologous systems when non-autonomous rice elements are introduced into yeast and Arabidopsis [10,18,19].

We found polymorphic putative full-size elements of twelve *Mutator*, three *Harbinger* and two *hAT* families. However, even among these large elements, most carried defective transposase ORFs which contained frameshifts or stop codons. The only exceptions were an insertion of the *hAT* element *DTA_Coraline* and several members of the *Mutator* family *DTM_MK*. Here, we found multiple copies that contain intact transposase ORFs. The *DTM_MK* family is particularly interesting because it illustrates how TEs can diverge into multiple sub-families. The *DTM_MK* family consists of multiple large elements that each contains a unique pattern of internal deletions of additionally acquired sequence fragments. Furthermore, it contains a sub-population of six small deletion derivatives that obviously originated from a single deletion event since they all have a very similar structure. These elements may represent the first steps in the evolution of a population of non-autonomous TEs.

DNA transposon excisions have a large potential to shape the genome

Of particular interest to us was a broad assessment of what types of footprints DNA transposons produce. We found that TE excisions can produce very complex patterns. Previous studies already suggested that excisions may produce a variety of outcomes and that the perfect footprint (that is, the precisely duplicated target site) might actually be rare [3]. Furthermore, it was shown

that excisions may lead to large deletions and/or insertions of copies of foreign DNA fragments when 'filler' DNA is inserted in the process of DSB repair [3,16,33]. Our data indeed show that perfect excisions which leave exactly two copies of the TSD are extremely rare, as only 8 out of 94 excisions showed this pattern. In all other cases, excisions lead to the deletion and/or introduction of foreign DNA fragments. Our large dataset allowed us to quantify that on average, 18 bp of the flanking region are deleted while 13 bp of the new sequence are introduced at the excision site. These numbers do not include the most extreme case wherein an excision apparently went along with the deletion of a 2.4 kb fragment. Furthermore, our dataset does not include possible cases where large segments on both sides of the element were deleted upon excision (such events would be indistinguishable from random deletions that by chance removed a large segment containing the TE). Also data from insertion/excision ratios of some superfamilies suggest that many excisions may have 'catastrophic' outcomes (see below). Thus, we conclude that excisions of DNA transposons are a major driving force in genome evolution as they can cause relatively large-scale rearrangements such as deletions and integrations of new sequences surrounding the excision site.

Why do *Harbinger* and *Mutator* elements show more insertions than expected?

The current model of proliferation during DNA replication postulates that one would find a ratio of insertions to excisions that lies somewhere between 1:1 and 2:1 (see Additional file 5: Figure S3 and Additional file 6: Figure S4) when comparing two closely related genomes. Interestingly, in all DNA transposon superfamilies, we found insertion/excision ratios higher than 2:1 in both species. Only for the *Mariner (DTT)* superfamily did we find a ratio of insertions to excisions that was only slightly above 2:1 in both *O. sativa* and *O. glaberrima*. In contrast, the insertion/excision ratios of the *Harbinger (DTH)* and *Mutator (DTM)* superfamilies are clearly higher than 2:1 (that is, they show a much higher number of insertions than expected). The same is also probably true for *CACTA (DTC)* elements, but there, the sample size is smaller and the insertion/excision ratio does not significantly deviate from 2:1.

One explanation for the distorted ratio is that, for some reason, we can simply not see excisions in our sequence alignments. Buchmann *et al.* [3] suggested that some excisions go along with deletions of several kb of the flanking regions. Indeed, for example for *Harbinger* elements we identified the highest proportion of "unclear" events. These comprise large sequence gaps which we could not clearly classify as excisions because too much sequence was deleted or rearranged surrounding the element. Thus, our hypothesis is that *Harbinger* and *Mutator* (and possibly *CACTA*) elements frequently cause large rearrangements (mostly large deletions) upon excision, so that the orthologous regions of the two species cannot be aligned easily anymore. If such deletions are in the size range of 3 kb to 5 kb, it would undermine our initial mapping of homologous loci that was based on blast searches of 5 kb segments. Additionally, if the fitness of a gamete carrying the excision is reduced or even lethal, this would also contribute to raising the ratio above the 2:1. One possible reason for frequent large deletions could be the size of the elements, simply because excisions of large elements may be more difficult to repair. Indeed, *Mariner* elements are on average the smallest of all the elements studied, and there, we find an insertion/excision ratio to be the closest to 2:1. With increasing average size of elements, we also see an increasing insertions/excision ratio.

A second explanation why we find fewer excisions than expected is that the DSB is repaired by using the sister chromatid as a template *via* the SDSA mechanism [34] analogous to what happens during gene conversion. In this case, the excision would be undetectable because it was repaired perfectly with a copy of the sister chromatid that still contains the insertion. Such reversion of excision sites has been described in *Drosophila melanogaster* [35] and *Caenorhabditis elegans* [36]. However, it is not clear why this repair mechanism would preferably be used in certain superfamilies such as *Harbinger* and *Mutator*.

Finally, it is possible that many excisions are precise (that is, the TE and one target site is removed) and thus could not be distinguished from insertions. This could, for example, explain our findings of the high ratios of 4.4:1 for *O. sativa* and 8.2:1 for *O. glaberrima* in the *Harbinger* superfamily. However, previous studies produced conflicting results on the frequency of precise excisions. Yang *et al.* [19] described that 83% of approximately all 30 excisions were precise for the *Harbinger* element *mPing* when expressing it in *A. thaliana*. In contrast, Kikuchi *et al.*, who worked with the same element in rice anther cultures, stated that only one case out of approximately 70 excision sites showed the footprint of a precise excision [16]. Thus, it is possible that the frequency of precise excisions depends on the conditions under which the transposition occurs. Additionally, the frequency of precise excisions could also differ between TE superfamilies. Indeed, for Mariner elements, we found a ratio close to 2:1, indicating that we were able to distinguish insertions and excisions well.

Conclusions

We conclude that the activity of DNA transposons (particularly the excision process) is a major evolutionary force driving the generation of genetic diversity. Additionally,

our data indicate that some DNA transposon excisions might cause such large-scale rearrangements so that they cannot be detected anymore. It is therefore likely that our study still under-estimates the impact of DNA transposon excisions on genome evolution. However, it will require further and more detailed studies of these transposable elements in multiple species to conclusively answer this question.

Methods

Genome-wide sequence alignments

The genome of *O. sativa* was split into fragments of 5 kb. Each of these fragments were then used in BLASTN searches against the *O. glaberrima* genome to identify the orthologous regions. As a primary filter criteria, we considered only fragments in the same orientation on the same chromosome with an identity of at least 96%. Then, 12 kb of sequence from both species (5 kb fragment + 7 kb adjacent 3′ sequence to create an overlap with the following fragment) were excised for pair-wise alignment. Here, we used the EMBOSS (emboss.sourceforge.net) program Water which implements the Smith-Waterman algorithm. We used a gap opening penalty of 30 and gap extension penalty of 0.1 to obtain alignments that preferably contain fewer but larger gaps.

Each of these pairs was scanned for alignment quality. We included all sequences that were embedded between at least 200 continuous bases that could be aligned with more than 90% perfect matches. The corresponding positions in the *O. sativa* genome were determined, and the overlapping individual alignments were re-assembled into one global alignment per chromosome. The consistency of the global alignment with the original assembly of *O. sativa* was tested extensively by manual comparisons of positions of randomly chosen sequences in and across the breakpoints of the overlaps. The global alignments were scanned for insertions or deletions (InDels) larger than 50 bp. InDels only separated by less than 4 bp were considered as one event. Additionally, InDels that bordered to sequence gaps (stretches of Ns) were discarded.

The remaining InDels were scanned for homologies to Class II TIR-order transposons from our in-house database that is derived from the TREP database (http://wheat.pw.usda.gov/ITMI/Repeats/) with the following BLASTN parameters: minimum alignment size of 50 bp and identity of at least 70%. The InDels that could not be associated to known TEs were used as a query for an iterative BLASTN search against the whole genome in which the InDel was found. Sequences with at least 15 copies and a minimum identity of 85% were considered putative TEs. The top 15 hits were extracted from the genome including a few hundred bp of flanking sequences. These were aligned with ClustalW to determine the precise borders of the element and to generate a consensus sequence. Consensus sequences were curated manually and added to the repeat database. Like this, we were able to expand the existing dataset for rice repeats at TREP from 59 sequences to 235 sequences. All scripts were written in PERL and are available upon request.

The data for this analysis were retrieved from Wang *et al.* [27] for *O. glaberrima* and the International Rice Genome Sequencing Project (IRGSP) for *O. sativa Nipponbare cultivar* [20] (plantbiology.msu.edu/pub/data/), respectively. We retrieved the annotation of the *O. sativa* genome from the Rice Genome Annotation Project [29] (Version 6/plantbiology.msu.edu). We removed all entries that included the word 'transpos' in the description line as well as putative genes (which also mostly correspond to TE sequences) and mapped the remaining genes on our version (version 5) of the genome using GMAP [37] (research-pub.gene.com/gmap/). For the gene ontology analysis, we used the online platform 'Rice Oligonucleotide Array Database' [38] (ricearray.org/analysis/) at default settings. We included the genes found to be affected by active TEs in exons, introns, 1,000 bp upstream, or 500 bp downstream of the CDS to check if they are often involved in certain biological processes disproportionately.

Estimate of the number of sequence gaps caused by DNA transposons

To assess the different assembly qualities, we first counted all Ns in both genomes. With a total of 93,930 Ns, the *O. sativa* assembly contains a low number of sequence gaps. In contrast, in *O. glaberrima*, we identified 20,080 gaps consisting of more than 50 Ns (total N count, 12,768,901). To study the cause of these sequence gaps, we extracted 500 bp up and downstream of these regions and identified the orthologous position in *O. sativa.* We identified 7,301 cases (4,413,818 Ns), where both flanking sequences mapped within 10 kb from each other in the same orientation (blast hits with a minimum of 400 bp length and 95% identity). We then screened the segment in *O. sativa* that corresponds to the gap in *O. glaberrima* for TE sequence. Of these orthologous loci, 25.6% (1,871 cases) showed homology to TIR DNA transposons. From this number, we extrapolated that proximately 5,150 sequence gaps in the *O. glaberrima* genome correspond to TIR DNA transposon sequences.

In the alignment of the two genomes, we identified 1,745 insertions (that is, additional sequence) in *O. glaberrima* larger than 50 bp which consist of more than 80% Ns. Assuming that about 25% of these loci correspond to DNA transposons, we expect 447 additional DNA transposon-related polymorphisms in *O. glaberrima*.

Additional data files

The following additional data are available with the online version of this paper. Additional file 1 is an illustration of

the different DSB repair mechanisms. Additional file 2 is a table listing all annotated DNA transposons in the genomes of *O. sativa* and *O. glaberrima*. Additional file 3 is a figure explaining different mechanisms that lead to InDels. Additional file 4 is a table listing all described TE polymorphisms. Additional file 5 is a figure explaining the inheritance of insertion and excision patterns of DNA transposons. Additional file 6 is a figure explaining that the differences in the ratio of insertions and excisions is an indicator for differential TE activity between species.

Additional files

Additional file 1: Figure S1. Overview of the different mechanisms of DSB repair following DNA transposon excision. A.) Non-homologous end joining (NHEJ) which leads to perfect excisions. B.) Single stranded annealing (SSA) which leads to a deletion of adjacent sequences. C.) Synthesis-dependent strand annealing (SDSA) which can lead to introduction of 'filler' DNA segments. D.) A special case of SSA which leads to precise excisions.

Additional file 2: Table S1. Overview of TE abundance in *O. sativa* and *O. glaberrima*.

Additional file 3: Figure S2. Summary of causes for insertions in rice species. Many of the identified insertions showed homology with DNA transposons but were not caused directly by their activity (for example, partial deletions of TEs). Therefore, we divided the remaining insertions into three classes based on their presumed molecular mechanism as follows: (i) repeat slippage, (ii) partial deletion, and (iii) unknown. Repeat slippage happens if DNA polymerase loses its template while synthesizing the new strand during replication and then re-adopts at a similar template close by. We found 149 insertions in *O. glaberrima* and 51 in *O. sativa* which represent differences in the number of tandem repeats between the two species. Template lengths ranged from simple dinucleotides to more than 20 bp. In two cases, entire TEs served as templates for slippage, deleting several kb between two elements. In these cases, unequal homologous cross over (similar to the mechanism that produces solo LTRs of retrotransposons) could be an alternative interpretation. Another 68 insertions in *O. glaberrima* and 94 in *O. sativa* resulted from partial deletions of TEs. These were deletions of apparently random segments within or close to TEs. Finally, 35 insertions in *O. glaberrima* and 66 in *O. sativa* could not be clearly classified. These InDels are often larger than the average InDel. These include cases where it was not possible to deduce the original, ancient state because, for example, multiple TEs were nested in these positions. Also included here are cases where a TE was found in the middle if a large insertion. These could potentially represent excisions which went along with deletions of large segments of the flanking sequence.

Additional file 4: Table S2. Overview of all transpositions.

Additional file 5: Figure S3. Inheritance of transposon insertion/excision patterns. For this model we assume that all transposition effects are selectively neutral. It is commonly accepted that a mechanism of multiplication is for DNA transposons to excise during DNA replication and to reinsert in front of the replication fork. This leads to one daughter strand with one copy of the element (A-type gamete) and one with two copies (B-type gamete). If a large number of transposons are active in many different loci in a species (this may be spread out over many generations), the offspring genome will be a mosaic of loci derived from A- and B-type gametes. When comparing that genome to that of a closely related species, loci resulting from A-type gametes will identify an excision and an insertion, while loci resulting from A-type gametes will only identify insertions. Thus the observed overall ratio of insertions to excisions from a given transposon family will be 2:1.

Additional file 6: Figure S4. Detection of differences in transposon activity in different species. This model assumes that a given transposon family was present in many copies in the ancestor species. After species divergence, the transposon family is active at different levels in the two species (100 transpositions in one and 200 in the other species). As described in Additional file 1: Figure S1, A- and B-type gametes are passed on to offspring in a 1:1 ratio. In a cross-species comparison which identifies transposons (additional sequences) which are present in one but absent in the other species, insertions in one species and excisions in the other will be detected. If a transposon family had different levels of activity in the two species since their divergence, insertion/excision ratios will deviate from the 2:1 ratio.

Abbreviations

TE: Transposable element; TIR: Terminal inverted repeat; MITE: Miniature inverted transposable element; TSD: Target site duplication; ORF: Open reading frame; DSB: Double-strand break; NHEJ: Non-homologous end joining; MMEJ: Microhomology-mediated end joining; SSA: Single strand annealing; SDSA: Synthesis-dependent strand annealing; InDel: Insertion or deletion.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SR performed the TE annotation and analysis and wrote the paper. TW designed the study and wrote the paper. Both authors have read and approved the final version of the paper.

Acknowledgements

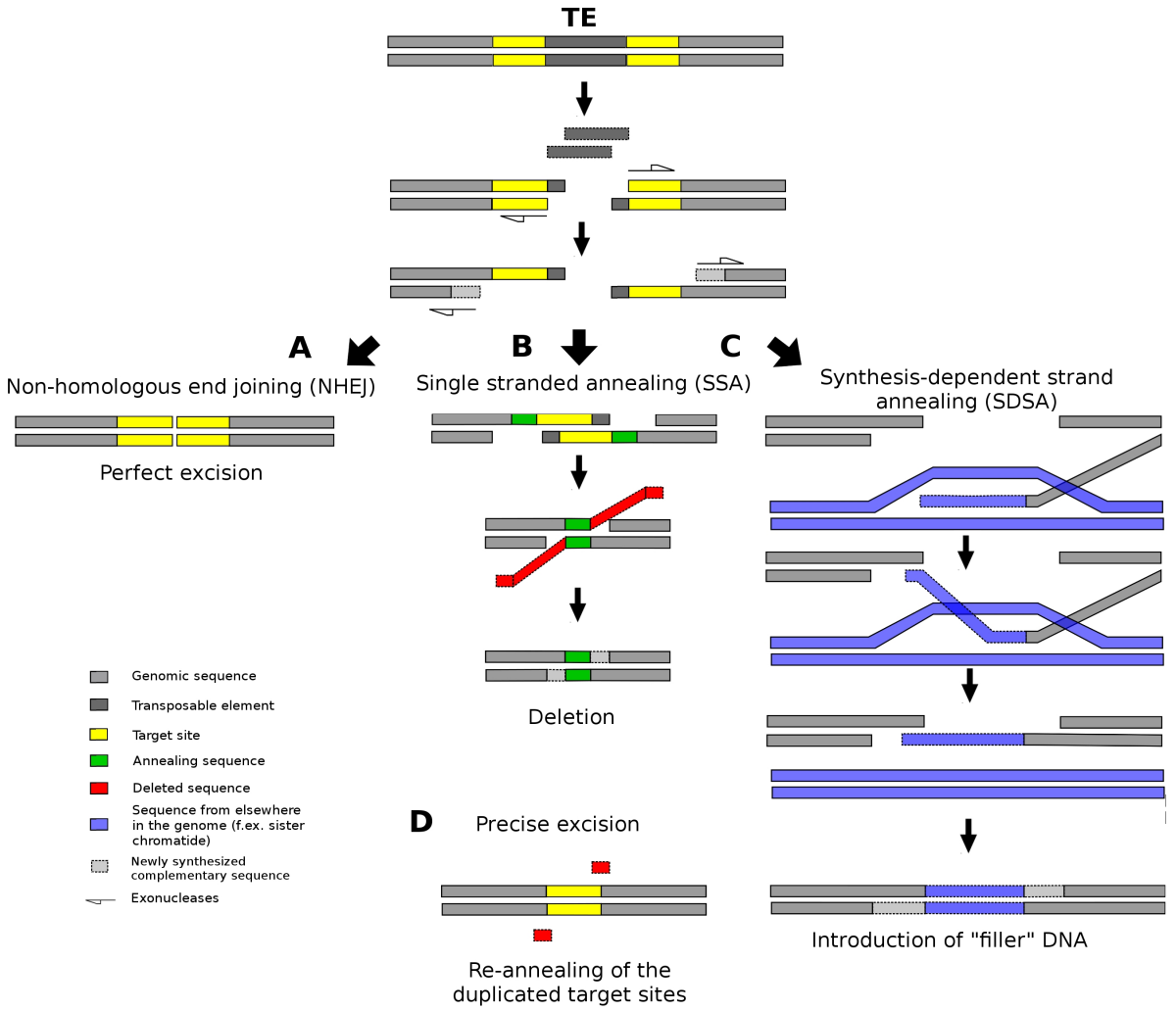
We would like to thank R. A. Wing for granting early access to the *O. glaberrima* data and G. Treier for support in the statistical analysis. This study was supported by the Swiss National Foundation grant # 31003A_138505/1.

Received: 19 January 2015 Accepted: 16 April 2015
Published online: 28 April 2015

References

1. Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet. 2007;41:331.
2. Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, et al. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. Genome Res. 2006;16:1262–9.
3. Buchmann JP, Matsumoto T, Stein N, Keller B, Wicker T. Inter-species sequence comparison of Brachypodium reveals how transposon activity corrodes genome colinearity. Plant J. 2012;71(4):550–63.
4. Wicker T, Buchmann JP, Keller B. Patching gaps in plant genomes results in gene movement and erosion of colinearity. Genome Res. 2010;20:1229–37.
5. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8(12):973–82.
6. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. The paleontology of intergene retrotransposons of maize. Nat Genet. 1998;20:43–5.
7. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, et al. The Sorghum bicolor genome and the diversification of grasses. Nature. 2009;457:551–6.
8. The International Brachypodium Initiative. Genome sequencing and analysis of the model grass Brachypodium distachyon. Nature. 2010;463:763–8.
9. Wicker T, Guyot R, Yahiaoui N, Keller B. CACTA transposons in Triticeae. A diverse family of high-copy repetitive elements. Plant Physiologist. 2003;132:52–63.
10. Yang G, Holligan Nagel D, Feschotte C, Hancock CN, Wessler SR. Tuned for transposition: molecular determinants underlying the hyperactivity of a Stowaway MITE. Science. 2009;325:1391–4.
11. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
12. Sabot F, Picault N, El-Baidouri M, Llauro C, Chaparro C, Piegu B, et al. Transpositional landscape of the rice genome revealed by paired-end mapping of high-throughput re-sequencing data. Plant J. 2011;66(2):241–6.
13. Bureau TE, Wessler SR. Stowaway: a new family of inverted-repeat elements associated with genes of both monocotyledonous and dicotyledonous plants. Plant Cell. 1994;6:907–16.

14. Bureau TE, Ronald PC, Wessler SR. A computer-based systematic survey reveals the predominance of small inverted-repeat elements in wild-type rice genes. Proc Natl Acad Sci. 1996;93:8524–9.
15. Nakazaki T, Okumoto Y, Horibata A, Yamahira S, Teraishi M, Nishida H, et al. Mobilization of a transposon in the rice genome. Nature. 2003;421:170–2.
16. Kikuchi K, Terauchi K, Wada M, Hirano HY. The plant MITE *mPing* is mobilized in anther culture. Nature. 2003;421:167–70.
17. Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR, et al. An active DNA transposon family in rice. Nature. 2003;421:163–7.
18. Yang G, Weil CF, Wessler SR. A rice Tc1/mariner-like element transposes in yeast. Plant Cell. 2006;18:2469–78.
19. Yang G, Zhang F, Hancock CN, Wessler SR. Transposition of the rice miniature inverted repeat transposable elemten *mPing* in *Arabidopsis thaliana*. Proc Natl Acad Sci U S A. 2007;104(26):10962–7.
20. Fujino K, Sekiguchi H, Kiguchi T. Identification of an active transposon in intact rice plants. Mol Gen Genomics. 2005;273:150–7.
21. Moon S, Jung KH, Lee DE, Jiang WZ, Koh HJ, Heu MH, et al. Identification of active transposon *dTok*, a member of the *hAT* family, in rice. Plant Cell Physiologist. 2006;47(11):1473–83.
22. Richardson JM, Colloms SD, Finnegan DJ, Walkinshaw MD. Molecular architecture of the Mos1 paired-end complex: the structural basis of DNA transposition in a eukaryote. Cell. 2009;138(6):1096–108.
23. Symington LS, Gautier J. Double-strand break end resection and repair pathway choice. Annu Rev Genet. 2011;45:247–71.
24. Edlinger B, Schlögelhofer P. Have a break: determinants of meiotic DNA double strand break (DSB) formation and processing in plants. J Exp Bot. 2011;62(5):1545–63.
25. Vu GTH, Cao HX, Watanabe K, Hensel G, Blattner FR, Kumlehn J. Schubert I: repair of site-specific DNA double-strand breaks in barley occurs via diverse pathways primarily involving the sister chromatid. Plant Cell. 2014;26(5):2156–67.
26. Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet. 2009;10:285–311.
27. Wang M, Yu Y, Haberer G, Marri PR, Fan C, Goicoechea JL, et al. The genome sequence of African rice (Oryza glaberrima) and evidence for independent domestication. Nat Genet. 2014;46:982–8.
28. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. Nature. 2005;436:793–800.
29. Ouyang S, Zhu W, Hamilton J, Haining L, Campbell M, Childs K, et al. The TIGR Rice Genome Annotation Resource: improvements and new features. Nucleic Acids Res. 2007;35:D883–7.
30. Wicker T, Keller B. Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. Genome Res. 2007;17:1072–81.
31. Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH. Genome evolution of wild barley (Hordeum spontaneum) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence. Proc Natl Acad Sci. 2000;97:6603–7.
32. Kimura M. On the probability of fixation of mutant genes in a population. Genetics. 1962;47:713–9.
33. Robert V, Bessereau JL. Targeted engineering of the Caenorhabditis elegans genome following Mos1-triggered chromosomal breaks. EMBO J. 2007;26:170–83.
34. Hartlerode AJ, Scully R. Mechanisms of double-strand break repair in somatic mammalian cells. Biochem J. 2009;423:157–68.
35. Engels WR, Johnson-Schlitz DM, Eggleston WB, Sved J. High-frequency P element loss in Drosophila is homolog dependent. Cell. 1990;62:515–25.
36. Plasterk RH. The origin of footprints of the Tc1 transposon of Caenorhabditis elegans. EMBO J. 1991;10:1919–25.
37. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21:1859–75.
38. Cao P, Jung KH, Choi D, Hwang D, Zhu J, Ronald PC. The Rice Oligonucleotide Array Database: an atlas of rice gene expression. Rice. 2012;5:17.



Additional Figure S1. Overview of the different mechanisms of DSB repair following DNA transposon excision. A.) Non-homologous end joining (NHEJ) which leads to perfect excisions. B.) Single stranded annealing (SSA) which leads to a deletion of adjacent sequences. C.) Synthesis-dependent strand annealing (SDSA) which can lead to introduction of “filler” DNA segments. D.) A special case of SSA which leads to precise

excisions.

-

Table S1: Overview of TE abundance in *O. sativa* and *O. glaberrima*

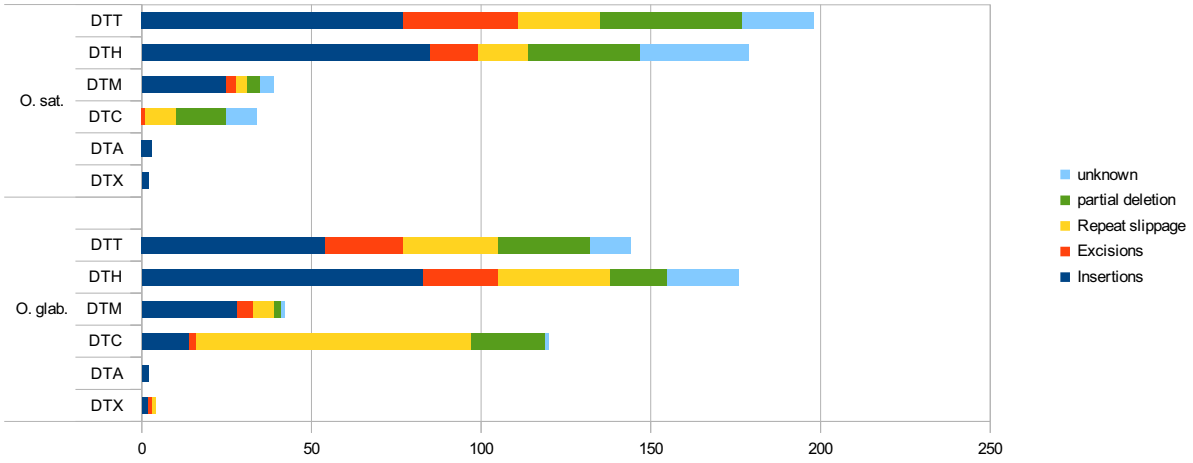
TE family	Copy number	Basepairs occupied	TE family	Copy number	Basepairs occupied
Mariner (DTT)			***Mariner (DTT)***		
DTT_SA	1875	220102	*DTT_SA*	1648	181016
DTT_SAA	214	30225	*DTT_SAA*	195	27307
DTT_SAB	79	5532	*DTT_SAB*	88	6046
DTT_SAC	115	10490	*DTT_SAC*	106	9467
DTT_SAD	1123	227077	*DTT_SAD*	1017	204483
DTT_SAF	699	87565	*DTT_SAF*	556	64707
DTT_SB	3995	487174	*DTT_SB*	3624	436358
DTT_SC	684	189223	*DTT_SC*	587	158365
DTT_SD	411	80172	*DTT_SD*	376	70226
DTT_SE	905	208112	*DTT_SE*	811	176028
DTT_SG	2403	461442	*DTT_SG*	2057	369602
DTT_SH	2060	347765	*DTT_SH*	1717	276194
DTT_SI	1063	213576	*DTT_SI*	834	155155
DTT_SJ	2354	247948	*DTT_SJ*	1809	184912
DTT_SK	496	101186	*DTT_SK*	423	84172
DTT_SM	2066	311754	*DTT_SM*	1842	272368
DTT_SN	241	53984	*DTT_SN*	223	47322
DTT_SQ	362	101019	*DTT_SQ*	332	91256
DTT_SR	25	1466	*DTT_SR*	22	1405
DTT_SS	803	151830	*DTT_SS*	674	126381
DTT_ST	612	57018	*DTT_ST*	528	50283
DTT_SU	148	31354	*DTT_SU*	134	26756
DTT_SV	173	14586	*DTT_SV*	140	11312
DTT_SW	978	177161	*DTT_SW*	892	160136
DTT_SX	304	50554	*DTT_SX*	285	44965
Harbinger (DTH)			***Harbinger (DTH)***		
DTH_Baba	52	14208	*DTH_Baba*	53	18994
DTH_Blip	68	51555	*DTH_Blip_A*	55	24292
DTH_OsKong	57	25927	*DTH_Kong*	62	32367
DTH_Pong	19	34467	*DTH_Pong*	9	861
DTH_TA	467	113832	*DTH_TA*	464	111840
DTH_TAA	288	68227	*DTH_TAA*	242	56944
DTH_TAB	143	26155	*DTH_TAB*	141	24325
DTH_TAC	48	6988	*DTH_TAC*	34	4959
DTH_TAD	1078	131160	*DTH_TAD*	877	104864
DTH_TAE	3022	332569	*DTH_TAE*	2426	254810
DTH_TAF	22	1733	*DTH_TAF*	17	1268
DTH_TAG	45	43781	*DTH_TAG*	25	15312
DTH_TAH	19	7408	*DTH_TAH*	24	34369
DTH_TAI	3028	601035	*DTH_TAI*	2764	539866
DTH_TAJ	179	29184	*DTH_TAJ*	192	32186
DTH_TAK	2	172	*DTH_TAK*	3	343
DTH_TAL	793	88893	*DTH_TAL*	672	73764
DTH_TAO	20	3591	*DTH_TAO*	19	3790
DTH_TAP	14	3840	*DTH_TAP*	11	2492
DTH_TAS	146	20824	*DTH_TAS*	129	18983
DTH_TAU	161	29966	*DTH_TAU*	159	29709
DTH_TB	109	13418	*DTH_TB*	100	13224

-

DTH_TC	1479	396387	*DTH_TC*	1202	301632
DTH_TD	131	9861	*DTH_TD*	131	9989
DTH_TE	510	123104	*DTH_TE*	457	110322
DTH_TF	1919	510105	*DTH_TF*	1745	452018
DTH_TG	937	165320	*DTH_TG*	796	135686
DTH_TI	93	8782	*DTH_TI*	89	8309
DTH_TO	3216	649618	*DTH_TO*	2955	595783
DTH_TR	581	148381	*DTH_TR*	399	99413
DTH_TS	1604	382053	*DTH_TS*	1183	264822
DTH_TT	505	43261	*DTH_TT*	389	33702
DTH_TU	103	11833	*DTH_TU*	93	11767
DTH_TV	224	34958	*DTH_TV*	207	32864
DTH_TW	50	9428	*DTH_TW*	83	18463
DTH_TY	152	26244	*DTH_TY*	102	16642
DTH_TZ	147	69196	*DTH_TZ*	133	56429
DTH_XAB	240	31875	*DTH_XAB*	228	30750
DTH_TX	294	92639	*DTH_TX*	261	76059
Mutator (DTM)			***Mutator (DTM)***		
DTM_HA	63	20859	*DTM_HA*	71	22988
DTM_MA	302	65813	*DTM_MA*	336	63054
DTM_MAA	212	60199	*DTM_MAA*	155	37006
DTM_MAB	86	11320	*DTM_MAB*	79	11603
DTM_MAC	308	56505	*DTM_MAC*	263	47824
DTM_MAD	2	1460	*DTM_MAD*	1	725
DTM_MAE	110	39384	*DTM_MAE*	79	27487
DTM_MAF	1408	234586	*DTM_MAF*	1232	207802
DTM_MAG	296	38061	*DTM_MAG*	265	32721
DTM_MB	202	94427	*DTM_MB*	196	93935
DTM_MC	251	138810	*DTM_MC*	255	89392
DTM_MD	152	47055	*DTM_MD*	143	42924
DTM_ME	622	110170	*DTM_ME*	542	93838
DTM_MF	64	27840	*DTM_MF*	78	32919
DTM_MG	74	29566	*DTM_MG*	46	26597
DTM_MH	78	11493	*DTM_MH*	74	10580
DTM_MJ	165	21874	*DTM_MJ*	124	15785
DTM_MK	672	1449063	*DTM_MK*	296	203120
DTM_MN	903	111517	*DTM_MN*	785	89989
DTM_MP	200	54956	*DTM_MP*	161	40220
DTM_MQ	118	48022	*DTM_MQ*	74	25182
DTM_MR	113	20488	*DTM_MR*	87	14383
DTM_MS	4	2385	*DTM_MS*	8	6766
DTM_MT	5	1286	*DTM_MT*	12	2522
DTM_MU	131	26646	*DTM_MU*	126	20587
DTM_MX	293	30976	*DTM_MX*	269	27413
DTM_MY	37	6092	*DTM_MY*	29	4729
DTM_MZ	123	13345	*DTM_MZ*	113	13898
DTM_XB	326	44185	*DTM_XB*	305	40352
CACTA (DTC)			***CACTA (DTC)***		
DTC_Alix	784	393380	*DTC_Alix*	533	176935
DTC_Baldur	21	30015	*DTC_Baldur*	29	20414
DTC_Benito	148	97417	*DTC_Benito*	140	81445
DTC_CA	188	44182	*DTC_CA*	155	39953
DTC_Calvin	4868	5511869	*DTC_Calvin*	3242	2099607
DTC_Carson	217	38003	*DTC_Carson*	193	30451

-

DTC_Dorian	369	462820	*DTC_Dorian*	275	185957
DTC_Eric	910	985533	*DTC_Eric*	818	764744
DTC_Grover	622	899717	*DTC_Grover*	406	358852
DTC_Isidor	285	39942	*DTC_Isidor*	262	23194
DTC_Janus	63	80190	*DTC_Janus*	55	47559
DTC_Radon	674	200825	*DTC_Radon*	313	93148
DTC_Rufus	22	43900	*DTC_Rufus*	22	14399
DTC_Sandro	67	36126	*DTC_Sandro*	91	77862
DTC_Seamus	20	28388	*DTC_Seamus*	14	10328
DTC_Sherman	137	60215	*DTC_Sherman*	125	47558
DTC_Storm	209	51912	*DTC_Storm*	198	53153
hAT (DTA)			***hAT (DTA)***		
DTA_Coraline	6	15292	*DTA_Coraline*	0	0
DTA_HA	16	5391	*DTA_HA*	7	3248
DTA_HB	192	25657	*DTA_HB*	172	24116
DTA_HC	122	19064	*DTA_HC*	108	16330
DTA_HD	482	112939	*DTA_HD*	379	71570
DTA_HE	244	22866	*DTA_HE*	219	20097
DTA_HG	32	4265	*DTA_HG*	37	7501
DTA_HI	22	3841	*DTA_HI*	12	1654
DTA_HJ	141	32099	*DTA_HJ*	118	26180
DTA_HK	92	45528	*DTA_HK*	59	16344
DTA_HL	73	8737	*DTA_HL*	49	5628
DTA_MI	146	20294	*DTA_MI*	162	23166



Additional Figure S2. Summary of causes for insertions in rice species. Many of the identified insertions showed homology with DNA transposons but were not caused directly by their activity (e.g. partial deletions of TEs). Therefore, were divided the remaining insertions in three classes based on their presumed molecular mechanism: (i) repeat slippage, (ii) partial deletion and (iii) unknown. Repeat slippage happens if DNA polymerase loses its template while synthesizing the new strand during replication and then re-adopts at a similar template close by. We found 149 insertion in *O. glaberrima* and 51 in the *O. sativa* which represent differences in the number of tandem repeats between the two species. Template lengths ranged from simple dinucleotides to more than 20 bp. In two cases, entire TEs served as templates for slippage, deleting several kb between two elements. In these cases, unequal homologous crossing-over (similar to the mechanism that produces solo LTRs of retrotransposons (ref)) could be an alternative interpretation. Another 68 insertions in *O. glaberrima* and 94 in *O. sativa*, resulted from partial deletions of TEs. These were deletions of apparently random segments within or close by TEs. Finally, 35 insertions in *O. glaberrima* and 66 in *O. sativa* could not be clearly classified. These indels are often larger than the average. These include cases where it was not possible to deduce the original, ancient state because, for example, multiple TEs were nested in these positions. Also included here are cases, where a TE was found in the middle if a large insertion. These could potentially represent excisions which went along with deletions of large segments of flanking sequence.

-

Table S2: Overview of all transpositions

ID	Excision / Insertion	TE family	Species	Chr.	Start *O. Sativa*	Start *O. glaberrima*	Target Site	bp deleted	bp „Filler“
1	E	*DTH_TS*	*O. sativa*	1	11361252	7664921	TAA	6	0
2	E	*DTM_MK*	*O. sativa*	1	12600385	9218754	CACCTCTTC / TCACCGTTCT	0	0
3	E	*DTH_TS*	*O. sativa*	1	12902570	9478733	TGA	0	13
4	E	*DTT_SAF*	*O. sativa*	1	14800923	10630380	TA	0	7
5	E	*DTT_SE*	*O. sativa*	1	2143974	1552674	TA	0	0
6	E	*DTT_SAF*	*O. sativa*	1	23428568	16010989	TA	17	16
7	E	*DTT_SI*	*O. sativa*	1	23745110	16234979	TA	5	0
8	E	*DTT_SH*	*O. sativa*	1	25116785	17535109	TA	0	1
9	E	*DTH_TR*	*O. sativa*	1	28103748	19840941	TAA	226	0
10	E	*DTT_SE*	*O. sativa*	1	28983077	20649311	TA	0	3
11	E	*DTT_SA*	*O. sativa*	1	29919153	21479533	TA	15	0
12	E	*DTM_MA*	*O. sativa*	1	33811302	24353085	ATGATAAAT	0	11
13	E	*DTH_TAS*	*O. sativa*	1	35000431	25182404	TTA	2	4
14	E	*DTC_Storm*	*O. sativa*	1	35288724	25434139	TAT	1	0
15	E	*DTH_TO*	*O. sativa*	1	41836087	31212136	TTA	10	2
16	E	*DTT_SA*	*O. sativa*	1	42564941	31717319	TA	0	1
17	E	*DTT_SG*	*O. sativa*	1	5150324	32601125	TA	5	0
18	E	*DTH_TF*	*O. sativa*	1	6649385	4959981	T(A/T)A	6	0
19	E	*DTT_SA*	*O. sativa*	2	11161618	9968245	TA	3	0
20	E	*DTT_SM*	*O. sativa*	2	1719100	1517049	TA	0	1
21	E	*DTT_SI*	*O. sativa*	2	21452665	17601179	TA	0	0
22	E	*DTH_TC*	*O. sativa*	2	21454670	17603056	TAA/TTC	5	0
23	E	*DTH_TA*	*O. sativa*	2	23428552	18987099	TT(T/A)	6	1
24	E	*DTT_SJ*	*O. sativa*	2	23974159	19453407	TA	3	2
25	E	*DTT_SG*	*O. sativa*	2	24072859	19545565	TA	23	161
26	E	*DTT_SH*	*O. sativa*	2	24169119	19638746	TA	4	3
27	E	*DTT_SJ*	*O. sativa*	2	27148252	21903384	TA	23	40
28	E	*DTT_SAF*	*O. sativa*	2	2725089	2463419	TA	0	2
29	E	*DTT_SC*	*O. sativa*	2	29194337	23380769	TA	0	12
30	E	*DTT_SG*	*O. sativa*	2	2963631	2706943	TA	14	0
31	E	*DTT_SJ*	*O. sativa*	2	29890827	23874746	TA	0	3
32	E	*DTT_SG*	*O. sativa*	2	33608046	27128831	TA	0	9
33	E	*DTH_TI*	*O. sativa*	2	34472772	27883970	T(G/A)G	12	2
34	E	*DTT_SJ*	*O. sativa*	2	35122138	28412790	TA	15	38
35	E	*DTH_TG*	*O. sativa*	2	400997	298568	TTA	3	0
36	E	*DTT_SM*	*O. sativa*	2	5051694	4634345	TA	0	1
37	E	*DTT_SJ*	*O. sativa*	2	5638072	5222271	TA	0	4
38	E	*DTT_SE*	*O. sativa*	2	6198541	5782264	TA	10	4
39	E	*DTT_SAF*	*O. sativa*	2	6453078	6059730	TA	121	6
40	E	*DTT_SI*	*O. sativa*	2	7026569	6419915	TA	0	1
41	E	*DTT_SC*	*O. sativa*	2	9717969	8789082	TA	0	1
42	E	*DTM_MH*	*O. sativa*	3	10435100	9646653	ATATATATATATGTAGAGAGA / TATATATAT	30	0
43	E	*DTT_SA*	*O. sativa*	3	11552786	10620135	TA	0	1
44	E	*DTT_SC*	*O. sativa*	3	12452418	11480541	TA	7	0
45	E	*DTH_TAA*	*O. sativa*	3	12884077	11897798	TAA	0	0
46	E	*DTT_SJ*	*O. sativa*	3	12918196	11942369	TA	16	4
47	I	*DTH_TG*	*O. sativa*	1	1029990	721289	TAA	0	0
48	I	*DTH_TAA*	*O. sativa*	1	10618137	8131027	TTA	0	0
49	I	*DTT_SJ*	*O. sativa*	1	1090465	786248	(T/C)A	0	0
50	I	*DTT_SJ*	*O. sativa*	1	1090465	786248	(T/C)A	0	0
51	I	*DTH_TC*	*O. sativa*	1	11079369	8420024	TAA	0	0
52	I	*DTH_TE*	*O. sativa*	1	11218178	7536702	TTA	0	0
53	I	*DTH_TAA*	*O. sativa*	1	11621340	32929750	TAA	0	0
54	I	*DTT_SG*	*O. sativa*	1	11675762	8504832	TA	0	0
55	I	*DTM_MK*	*O. sativa*	1	12105841	8797684	GAGCTGTCAA	0	0
56	I	*DTH_TS*	*O. sativa*	1	12350118	9044820	TTA	0	0
57	I	*DTT_SE*	*O. sativa*	1	12533001	9162528	TA	0	0
58	I	*DTM_MAA*	*O. sativa*	1	12664082	9283026	TTATTTTAA	0	0
59	I	*DTM_MAA*	*O. sativa*	1	12664082	9283026	TTATTTTAA	0	0
60	I	*DTH_TS*	*O. sativa*	1	12692397	9314114	CTC	0	0
61	I	*DTT_ST*	*O. sativa*	1	12714106	9326341	TA	0	0
62	I	*DTH_TY*	*O. sativa*	1	12864182	9437759	TTA	0	0
63	I	*DTT_SI*	*O. sativa*	1	13501578	9686190	TA	0	0
64	I	*DTM_MA*	*O. sativa*	1	14126284	10258650	GTCT(T/A)AACC	0	0
65	I	*DTH_TG*	*O. sativa*	1	14215209	10308991	TTA	0	0
66	I	*DTT_SAF*	*O. sativa*	1	14798576	10627374	TA	0	0
67	I	*DTC_Calvin*	*O. sativa*	1	15342434	10707383	TTA	0	0
68	I	*DTH_TS*	*O. sativa*	1	15345607	10709775	TAA	0	0
69	I	*DTC_Calvin*	*O. sativa*	1	15595678	10925172	TGG	0	0
70	I	*DTC_Alix*	*O. sativa*	1	15903567	11201038	GAA	0	0
71	I	*DTC_Benito*	*O. sativa*	1	17775468	12554804	CAC	0	0

-

72	I	*DTM_MK*	*O. sativa*	1	17931123	12708519	CGTGAATAGA	0	0
73	I	*DTH_TF*	*O. sativa*	1	18852422	13491709	TCA	0	0
74	I	*DTT_SAF*	*O. sativa*	1	18924946	13542799	TA	0	0
75	I	*DTT_SW*	*O. sativa*	1	19311706	13788640	TA	0	0
76	I	*DTT_SA*	*O. sativa*	1	19693153	14193910	TA	0	0
77	I	*DTH_Blip_A*	*O. sativa*	1	20703120	14567773	TAA	0	0
78	I	*DTH_TAI*	*O. sativa*	1	21066090	14732026	TAA	0	0
79	I	*DTT_SJ*	*O. sativa*	1	2139659	1548391	TA	0	0
80	I	*DTM_MA*	*O. sativa*	1	21500172	15074350	TACGGAGAT	0	0
81	I	*DTT_SA*	*O. sativa*	1	21567491	15121703	TA	0	0
82	I	*DTA_HJ*	*O. sativa*	1	21713076	15206684	AAATAATA	0	0
83	I	*DTH_TC*	*O. sativa*	1	21716777	15210337	TTA	0	0
84	I	*DTT_SJ*	*O. sativa*	1	21962983	15450955	TA	0	0
85	I	*DTH_TR*	*O. sativa*	1	22011604	15512019	TCA	0	0
86	I	*DTH_TE*	*O. sativa*	1	22023762	15526376	TTC	0	0
87	I	*DTT_SG*	*O. sativa*	1	22242282	15723374	TA	0	0
88	I	*DTH_TC*	*O. sativa*	1	23690198	16180807	TTA	0	0
89	I	*DTC_Calvin*	*O. sativa*	1	23709314	16195468	CTT	0	0
90	I	*DTT_SAF*	*O. sativa*	1	23993262	16435693	TA	0	0
91	I	*DTT_SG*	*O. sativa*	1	24012962	16451074	TA	0	0
92	I	*DTT_SJ*	*O. sativa*	1	24046421	16482148	TA	0	0
93	I	*DTC_Calvin*	*O. sativa*	1	24046828	16482349	AGG	0	0
94	I	*DTH_TC*	*O. sativa*	1	24139167	16578059	TAA	0	0
95	I	*DTH_TAO*	*O. sativa*	1	24409463	16831028	TTA	0	0
96	I	*DTH_TE*	*O. sativa*	1	24511445	16932491	TAA	0	0
97	I	*DTH_TR*	*O. sativa*	1	24777404	17209381	TTA	0	0
98	I	*DTH_TAA*	*O. sativa*	1	24832193	17264284	TTA	0	0
99	I	*DTC_Calvin*	*O. sativa*	1	25208673	17622358	ATT	0	0
100	I	*DTH_TS*	*O. sativa*	1	25478459	17886088	TTA	0	0
101	I	*DTH_OsKong*	*O. sativa*	1	25587961	17971265	TTA	0	0
102	I	*DTT_SH*	*O. sativa*	1	25929275	18125239	TA	0	0
103	I	*DTH_TG*	*O. sativa*	1	25996960	18182003	TAA	0	0
104	I	*DTC_Calvin*	*O. sativa*	1	26149708	18304811	TAA	0	0
105	I	*DTT_SA*	*O. sativa*	1	26194275	18340826	TA	0	0
106	I	*DTM_MC*	*O. sativa*	1	26232617	18365233	TTATAAAAT	0	0
107	I	*DTH_TC*	*O. sativa*	1	26615893	18676645	TAA	0	0
108	I	*DTH_TR*	*O. sativa*	1	26793849	18849018	TAA	0	0
109	I	*DTM_MAB*	*O. sativa*	1	26800778	18857349	CCCAAAATA	0	0
110	I	*DTT_SJ*	*O. sativa*	1	270582	195165	TA	0	0
111	I	*DTC_Calvin*	*O. sativa*	1	26969062	18978096	GGA	0	0
112	I	*DTH_TR*	*O. sativa*	1	27007522	18995540	TAA	0	0
113	I	*DTC_Benito*	*O. sativa*	1	27056979	19052371	GTA	0	0
114	I	*DTH_TC*	*O. sativa*	1	27302063	19242698	TTA	0	0
115	I	*DTH_TY*	*O. sativa*	1	27432878	19345246	TAA	0	0
116	I	*DTH_TAP*	*O. sativa*	1	2753187	2075913	TTA	0	0
117	I	*DTT_SAF*	*O. sativa*	1	27810679	19645995	TA	0	0
118	I	*DTC_Calvin*	*O. sativa*	1	27844500	19675740	GTT	0	0
119	I	*DTM_MA*	*O. sativa*	1	28092931	19833547	CTTTTATTT	0	0
120	I	*DTC_Grover*	*O. sativa*	1	28175915	19900871	AAG	0	0
121	I	*DTH_TC*	*O. sativa*	1	2805867	2129304	TAA	0	0
122	I	*DTH_TAD*	*O. sativa*	1	2806533	2129625	TGA	0	0
123	I	*DTH_TR*	*O. sativa*	1	28682974	20360981	TTA	0	0
124	I	*DTT_SK*	*O. sativa*	1	28699841	20376114	TA	0	0
125	I	*DTH_TR*	*O. sativa*	1	28764004	20443867	TAA	12	0
126	I	*DTH_TC*	*O. sativa*	1	28818278	20498889	TAA	0	0
127	I	*DTT_SG*	*O. sativa*	1	2871106	2194949	TA	0	0
128	I	*DTT_SAF*	*O. sativa*	1	28939358	20609588	TA	0	0
129	I	*DTT_SG*	*O. sativa*	1	28969568	20640436	TA	0	0
130	I	*DTH_TAI*	*O. sativa*	1	29000318	20666546	TTA	0	0
131	I	*DTM_MK*	*O. sativa*	1	29025892	20701769	CACTCTGTT	0	0
132	I	*DTH_TAA*	*O. sativa*	1	29144938	20797095	TTA	0	0
133	I	*DTH_TR*	*O. sativa*	1	29536779	21095477	TAA	0	0
134	I	*DTH_TAA*	*O. sativa*	1	29539166	21097583	TTA	0	0
135	I	*DTA_HI*	*O. sativa*	1	29663519	21227687	ATTGTATT	0	0
136	I	*DTH_TAA*	*O. sativa*	1	29777774	21309387	TTA	0	0
137	I	*DTT_SH*	*O. sativa*	1	29898371	21457014	TA	0	0
138	I	*DTH_TO*	*O. sativa*	1	29900074	21458466	TAA	0	0
139	I	*DTT_SI*	*O. sativa*	1	30637091	21978760	TA	0	0
140	I	*DTT_SJ*	*O. sativa*	1	3070643	2354740	TA	0	0
141	I	*DTH_TS*	*O. sativa*	1	30827768	22046978	TCA	0	0
142	I	*DTT_SH*	*O. sativa*	1	3137939	2412828	TA	0	0
143	I	*DTT_SG*	*O. sativa*	1	31363808	22341475	TA	0	0
144	I	*DTM_MK*	*O. sativa*	1	31513352	22474800	TTAGTATTAT / TTAGTACTA	1	0
145	I	*DTH_TAC*	*O. sativa*	1	32072664	22852509	TTA	0	0
146	I	*DTT_SC*	*O. sativa*	1	32526213	23235177	TA	0	0
147	I	*DTH_TR*	*O. sativa*	1	32527477	23236148	TTA	0	0
148	I	*DTT_SX*	*O. sativa*	1	32747903	23426262	TA	0	0
149	I	*DTM_MN*	*O. sativa*	1	32770865	23449069	GCTACAGAA	0	0
150	I	*DTT_SG*	*O. sativa*	1	32773764	23451346	TA	0	0
151	I	*DTM_MU*	*O. sativa*	1	32793489	23471138	GAATTTGAA	0	0

-

152	I	*DTH_TAA*	*O. sativa*	1	32807445	23480472	TAT	0	0
153	I	*DTH_TO*	*O. sativa*	1	32809218	23481180	TTA	0	0
154	I	*DTH_TC*	*O. sativa*	1	32822430	23481287	TAA	0	0
155	I	*DTH_TS*	*O. sativa*	1	32824934	23483357	TTA	0	0
156	I	*DTH_TR*	*O. sativa*	1	33177825	23827330	TAA	0	0
157	I	*DTH_TR*	*O. sativa*	1	33412795	24043583	TAA	0	0
158	I	*DTH_TR*	*O. sativa*	1	33413061	24043849	TAA	0	0
159	I	*DTM_MAD*	*O. sativa*	1	33414482	24044744	TTTAAATTT / TTTTTTTAATTT	3	0
160	I	*DTT_SH*	*O. sativa*	1	33507535	24098832	TA	0	0
161	I	*DTM_MK*	*O. sativa*	1	33531341	24123419	TAGACCCGA	0	0
162	I	*DTC_Calvin*	*O. sativa*	1	33667361	24212843	CGC	0	0
163	I	*DTT_SA*	*O. sativa*	1	34275147	24687769	TA	0	0
164	I	*DTA_HK*	*O. sativa*	1	34506671	24875789	CTCAGGGC(T/C)	0	0
165	I	*DTH_TO*	*O. sativa*	1	34801275	24980553	TTA	0	0
166	I	*DTT_SM*	*O. sativa*	1	34813658	24993091	TA	0	0
167	I	*DTH_TAF*	*O. sativa*	1	3489803	2723711	TTA	0	0
168	I	*DTM_MP*	*O. sativa*	1	34978639	25162648	CGCGGTGCA	0	0
169	I	*DTT_SA*	*O. sativa*	1	34999123	25181361	TA	0	0
170	I	*DTM_MQ*	*O. sativa*	1	35106477	25274490	ACTAGCAGA	0	0
171	I	*DTH_TR*	*O. sativa*	1	35109128	25276535	TA(A/G)	0	0
172	I	*DTT_SC*	*O. sativa*	1	35110147	25277210	TA	0	0
173	I	*DTH_TR*	*O. sativa*	1	35194415	25351719	TAA	0	0
174	I	*DTH_TS*	*O. sativa*	1	35515368	25640062	TAA	0	0
175	I	*DTC_Grover*	*O. sativa*	1	35792566	25870869	TTT	0	0
176	I	*DTM_MK*	*O. sativa*	1	35922945	24813066	ACGGTTAGC	0	0
177	I	*DTH_TC*	*O. sativa*	1	36042918	26185229	TTA	0	0
178	I	*DTT_SJ*	*O. sativa*	1	3622426	22511772	TA	0	0
179	I	*DTA_Coraline*	*O. sativa*	1	3622629	22511975	CGGAAACC	0	0
180	I	*DTM_MA*	*O. sativa*	1	36319685	26371502	ATAAATGAG	0	0
181	I	*DTH_TC*	*O. sativa*	1	3637091	22489351	TCA	0	0
182	I	*DTM_MAE*	*O. sativa*	1	36837071	26870950	GACTCTATG / TAGACTCTATG	2	0
183	I	*DTH_TR*	*O. sativa*	1	37301352	27211559	TTA	0	0
184	I	*DTH_TC*	*O. sativa*	1	37803012	27664111	TTC	0	0
185	I	*DTH_TX*	*O. sativa*	1	38202775	28020955	GTT	0	0
186	I	*DTH_TC*	*O. sativa*	1	38306346	28121068	TCA	0	0
187	I	*DTH_TC*	*O. sativa*	1	38384786	28198855	TTA	0	0
188	I	*DTH_TC*	*O. sativa*	1	38488039	28312458	T(C/T)A	0	0
189	I	*DTM_MR*	*O. sativa*	1	39183038	28963448	ACAATATAA	0	0
190	I	*DTM_MA*	*O. sativa*	1	39235137	29002088	AACTTGATG	0	0
191	I	*DTT_SG*	*O. sativa*	1	393663	302172	TA	0	0
192	I	*DTH_Kong*	*O. sativa*	1	39536580	29286687	TTA	0	0
193	I	*DTH_TR*	*O. sativa*	1	394030	302302	TT(A/C)	0	0
194	I	*DTH_TF*	*O. sativa*	1	39751967	29485011	TTA	0	0
195	I	*DTM_MA*	*O. sativa*	1	40450026	30134455	GTGTATTTA	0	0
196	I	*DTH_TG*	*O. sativa*	1	40743012	30404581	TAA	0	0
197	I	*DTT_SI*	*O. sativa*	1	41058142	30493453	TA	0	0
198	I	*DTM_MA*	*O. sativa*	1	41635434	31073955	TTTATGCAG	0	0
199	I	*DTT_SI*	*O. sativa*	1	41861760	31239649	TA	0	0
200	I	*DTM_MU*	*O. sativa*	1	41872930	31250773	CATAAGTAA	0	0
201	I	*DTH_TS*	*O. sativa*	1	4180877	3397035	TAA	0	0
202	I	*DTT_SG*	*O. sativa*	1	42167895	31436994	TA	0	0
203	I	*DTH_TE*	*O. sativa*	1	42217169	31482342	TTA	0	0
204	I	*DTT_SJ*	*O. sativa*	1	42228825	31493868	TA	0	0
205	I	*DTH_TX*	*O. sativa*	1	42265420	31530204	TTA	0	0
206	I	*DTH_TO*	*O. sativa*	1	42371967	31623913	TTA	0	0
207	I	*DTH_TA*	*O. sativa*	1	42564757	31717272	TAA	0	0
208	I	*DTT_SC*	*O. sativa*	1	4248654	3451587	TA	0	0
209	I	*DTM_MA*	*O. sativa*	1	42968458	32086827	TTAATTAGA	0	0
210	I	*DTT_SG*	*O. sativa*	1	43049231	32152731	TA	0	0
211	I	*DTT_SA*	*O. sativa*	1	43071509	32176237	TA	0	0
212	I	*DTM_MB*	*O. sativa*	1	43494073	32515318	TAAGTATT(G/A)	0	0
213	I	*DTH_TG*	*O. sativa*	1	4576059	3708481	TTA	0	0
214	I	*DTT_SX*	*O. sativa*	1	4828828	3865470	TA	0	0
215	I	*DTM_MK*	*O. sativa*	1	4836302	3873930	GTCCTATAT	0	0
216	I	*DTH_TR*	*O. sativa*	1	5084440	4092388	TAA	0	0
217	I	*DTT_SA*	*O. sativa*	1	5148338	32599016	TA	0	0
218	I	*DTH_TS*	*O. sativa*	1	5270639	21558334	TTA	0	0
219	I	*DTT_SI*	*O. sativa*	1	5274465	21561821	TA	0	0
220	I	*DTH_TS*	*O. sativa*	1	5362817	4316810	TTA	0	0
221	I	*DTM_MK*	*O. sativa*	1	543859	433660	TAATAGGAG	0	0
222	I	*DTT_SG*	*O. sativa*	1	5525433	4466706	TA	0	0
223	I	*DTH_TF*	*O. sativa*	1	577920	473181	CAA	0	0
224	I	*DTT_SG*	*O. sativa*	1	612468	506197	TA	0	0
225	I	*DTH_TS*	*O. sativa*	1	6217023	4568566	TAA	0	0
226	I	*DTT_SG*	*O. sativa*	1	6282708	4634701	TA	0	0
227	I	*DTH_TAA*	*O. sativa*	1	6437811	4763951	TTA	0	0
228	I	*DTM_MK*	*O. sativa*	1	723601	593019	GTGCAAACG	0	0
229	I	*DTT_SJ*	*O. sativa*	1	7270743	5514266	TA	0	0
230	I	*DTT_SG*	*O. sativa*	1	7303595	5542142	TA	0	0

-

231	I	*DTH_TR*	*O. sativa*	1	7499952	5693290	TTA	0	0
232	I	*DTT_SI*	*O. sativa*	1	78791	33903	TA	0	0
233	I	*DTT_SJ*	*O. sativa*	1	8022493	6149028	TA	0	0
234	I	*DTT_SC*	*O. sativa*	1	8084430	6224065	TA	0	0
235	I	*DTH_TO*	*O. sativa*	1	8795709	6889216	TAA	0	0
236	I	*DTA_HC*	*O. sativa*	1	8966844	7025200	ATCAGAAC	0	0
237	I	*DTT_SH*	*O. sativa*	1	9105981	7143780	TA	0	0
238	I	*DTH_TB*	*O. sativa*	1	9690186	7307353	TAA	0	0
239	E	*DTH_TA*	*O. glaberrima*	1	10528347	8032425	TAA	4	85
240	E	*DTH_TG*	*O. glaberrima*	1	10934179	8286779	GGT/ATA	0	0
241	E	*DTT_SD*	*O. glaberrima*	1	12632720	9251948	TA	10	0
242	E	*DTH_TO*	*O. glaberrima*	1	1339109	961268	TAA	0	40
243	E	*DTC_Calvin*	*O. glaberrima*	1	15197276	10707364	TTA	0	33
244	E	*DTM_MAC*	*O. glaberrima*	1	17556651	12361801	AAAATTAAA	1	0
245	E	*DTT_SA*	*O. glaberrima*	1	20924040	14704564	TA	1	1
246	E	*DTT_SQ*	*O. glaberrima*	1	23493205	16063204	TA	5	6
247	E	*DTH_TC*	*O. glaberrima*	1	25566676	17945502	TTA	0	100
248	E	*DTM_MAA*	*O. glaberrima*	1	25594551	17977187	TATAATTAA	18	11
249	E	*DTH_TG*	*O. glaberrima*	1	26810539	18866867	TAA	0	13
250	E	*DTH_TAE*	*O. glaberrima*	1	269507	194331	TTA	0	1
251	E	*DTM_MK*	*O. glaberrima*	1	26894172	18917971	TCAGAGTTC	0	15
252	E	*DTH_TO*	*O. glaberrima*	1	27695981	19606401	TTA	0	0
253	E	*DTH_TG*	*O. glaberrima*	1	27698927	19608776	TA(C/A)	0	9
254	E	*DTT_SE*	*O. glaberrima*	1	27706971	19614664	TA	0	0
255	E	*DTT_SJ*	*O. glaberrima*	1	27807094	19642683	TA	0	2
256	E	*DTT_SS*	*O. glaberrima*	1	28400216	20070370	TA	18	0
257	E	*DTT_SJ*	*O. glaberrima*	1	28465845	20139126	TA	0	1
258	E	*DTT_SA*	*O. glaberrima*	1	28760719	20441417	TA	6	0
259	E	*DTH_TO*	*O. glaberrima*	1	29925102	21491078	TTA	5	0
260	E	*DTT_SAF*	*O. glaberrima*	1	304763	211339	TA	0	10
261	E	*DTH_TR*	*O. glaberrima*	1	32417805	23123249	TTA	5	0
262	E	*DTH_TG*	*O. glaberrima*	1	32844031	23502369	TTA	22	0
263	E	*DTH_TS*	*O. glaberrima*	1	32995743	23645250	CTT/AAT	0	0
264	E	*DTT_SC*	*O. glaberrima*	1	34727057	24894673	TA	2479	2
265	E	*DTM_MK*	*O. glaberrima*	1	37710749	27573608	CTTGGGCGG / GTTTCTAA	0	19
266	E	*DTT_SG*	*O. glaberrima*	1	3855089	3064188	TA	3	4
267	E	*DTH_TG*	*O. glaberrima*	1	38726913	28525853	TAA	1	1
268	E	*DTT_SI*	*O. glaberrima*	1	39091831	28887455	TA	0	16
269	E	*DTH_TAG*	*O. glaberrima*	1	39194889	28973586	TTA	0	0
270	E	*DTT_SG*	*O. glaberrima*	1	397579	22916371	TA	7	0
271	E	*DTH_TR*	*O. glaberrima*	1	3995117	3217237	TTA	19	3
272	E	*DTH_TS*	*O. glaberrima*	1	40162275	29861404	TAA	1	0
273	E	*DTT_SI*	*O. glaberrima*	1	40377109	30069044	TA	3	2
274	E	*DTH_TG*	*O. glaberrima*	1	41420991	30853691	TAA	1	0
275	E	*DTT_SG*	*O. glaberrima*	1	41435175	30867749	TA	5	14
276	E	*DTH_TO*	*O. glaberrima*	1	41475508	30910997	TAA	3	0
277	E	*DTT_SM*	*O. glaberrima*	1	41552023	31001343	TA	0	1
278	E	*DTH_TR*	*O. glaberrima*	1	41647894	31084472	TAA	94	0
279	E	*DTM_MK*	*O. glaberrima*	1	42159590	31430923	TTTCCAACT	12	44
280	E	*DTT_SA*	*O. glaberrima*	1	42290224	31555204	TA	0	0
281	E	*DTT_SJ*	*O. glaberrima*	1	4445243	3618838	TA	0	5
282	E	*DTT_SI*	*O. glaberrima*	1	7088336	5381933	TA	0	1
283	E	*DTT_SG*	*O. glaberrima*	1	856012	627122	TA	7	9
284	E	*DTM_MAG*	*O. glaberrima*	1	8527114	6669527	TTACTAGTA	15	7
285	E	*DTM_MK*	*O. glaberrima*	1	8801090	6894249	CCATCTATA	14	2
286	E	*DTT_SJ*	*O. glaberrima*	2	35123671	28414477	TA	54	123
287	I	*DTH_TO*	*O. glaberrima*	1	10704193	8215296	TTA	0	0
288	I	*DTT_ST*	*O. glaberrima*	1	10875441	8234201	TA	0	0
289	I	*DTH_TW*	*O. glaberrima*	1	10980969	8327093	T(G/A)A	0	0
290	I	*DTM_XB*	*O. glaberrima*	1	11654383	8483867	(T/A)TATTAATT	0	0
291	I	*DTA_MI*	*O. glaberrima*	1	11677212	8506567	TTAG(A/C)ATT	0	0
292	I	*DTT_SV*	*O. glaberrima*	1	12635710	9254649	TA	0	0
293	I	*DTH_TF*	*O. glaberrima*	1	12680243	9299302	TAA	0	0
294	I	*DTM_MZ*	*O. glaberrima*	1	14801139	10630682	TTTTTAAAA	0	0
295	I	*DTM_MA*	*O. glaberrima*	1	16513446	11400157	ATGTTTCAA	0	0
296	I	*DTH_TO*	*O. glaberrima*	1	17703359	12488522	TCA	0	0
297	I	*DTT_SA*	*O. glaberrima*	1	18924818	13542482	TA	0	0
298	I	*DTH_TO*	*O. glaberrima*	1	20306244	14356850	TAA	0	0
299	I	*DTH_TG*	*O. glaberrima*	1	20578842	14468760	TTA	0	0
300	I	*DTT_SJ*	*O. glaberrima*	1	2143553	1552015	TA	0	0
301	I	*DTT_SA*	*O. glaberrima*	1	21677622	15174678	TA	0	0
302	I	*DTT_SAF*	*O. glaberrima*	1	21968452	15456544	TA	0	0
303	I	*DTT_SG*	*O. glaberrima*	1	22039788	15542210	TA	0	0
304	I	*DTM_MD*	*O. glaberrima*	1	22043909	15547010	TTTTAAAAA	0	0
305	I	*DTH_TO*	*O. glaberrima*	1	23434952	16017266	TAA	0	0
306	I	*DTT_SG*	*O. glaberrima*	1	23744590	16234223	TA	0	0
307	I	*DTH_TC*	*O. glaberrima*	1	24002093	16440426	TTA	0	0
308	I	*DTH_TG*	*O. glaberrima*	1	24019574	16458436	TAA	0	0
309	I	*DTM_MC*	*O. glaberrima*	1	24091648	16527920	ATTCTTCTT	0	0

-

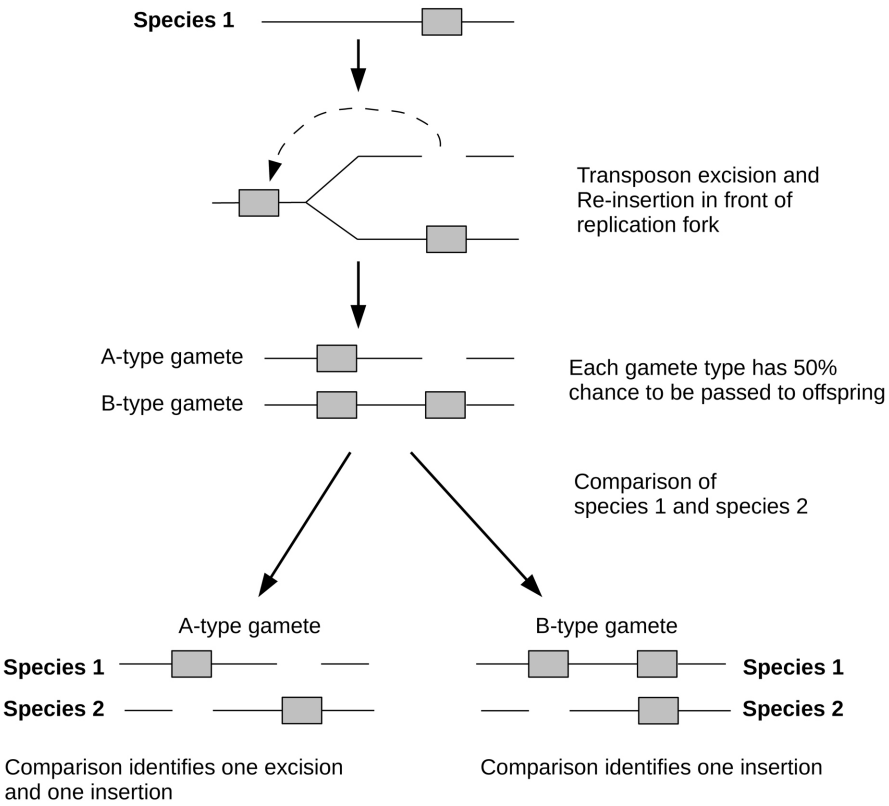
310	I	*DTT_ST*	*O. glaberrima*	1	24171930	16611975	TA	0	0
311	I	*DTM_MA*	*O. glaberrima*	1	24692360	17122961	TTATCAGTA	0	0
312	I	*DTH_TC*	*O. glaberrima*	1	25100147	17523297	TAA	0	0
313	I	*DTT_SJ*	*O. glaberrima*	1	25194761	17608772	TA	0	0
314	I	*DTM_MA*	*O. glaberrima*	1	25347445	17764949	AAGAAGCAG	0	0
315	I	*DTM_MA*	*O. glaberrima*	1	25600673	17983050	GTAGTTAAC	0	0
316	I	*DTM_MAF*	*O. glaberrima*	1	25945747	18135717	AGCTTTCAT	0	0
317	I	*DTA_HL*	*O. glaberrima*	1	25945579	18135549	CTTG(C/T)GTC	0	0
318	I	*DTM_MA*	*O. glaberrima*	1	2609560	1947187	TTATAGTAG	0	0
319	I	*DTH_TAA*	*O. glaberrima*	1	26160512	18314521	TTA	0	0
320	I	*DTH_TS*	*O. glaberrima*	1	26386988	18496175	TAA	0	0
321	I	*DTH_TW*	*O. glaberrima*	1	27575049	19486786	TAA	0	0
322	I	*DTT_SC*	*O. glaberrima*	1	28157347	19879800	TA	0	0
323	I	*DTH_TAA*	*O. glaberrima*	1	28395887	20065798	TTA	0	0
324	I	*DTH_TO*	*O. glaberrima*	1	28481996	20155402	TTA	0	0
325	I	*DTT_SJ*	*O. glaberrima*	1	28705765	20384749	TA	0	0
326	I	*DTH_TF*	*O. glaberrima*	1	28708553	20388052	AAG	0	0
327	I	*DTT_SAF*	*O. glaberrima*	1	2874183	2197794	TA	0	0
328	I	*DTT_SE*	*O. glaberrima*	1	29263895	20875104	TA	0	0
329	I	*DTT_SV*	*O. glaberrima*	1	29314897	20927422	(G/T)A	3	1
330	I	*DTM_MA*	*O. glaberrima*	1	29432840	20999747	CAGAATCAA	0	0
331	I	*DTT_SW*	*O. glaberrima*	1	29912091	21473105	TA	0	0
332	I	*DTH_TS*	*O. glaberrima*	1	30003221	21627040	T(A/G)A	0	0
333	I	*DTH_OsKong*	*O. glaberrima*	1	309517	215967	ATA	0	0
334	I	*DTH_TO*	*O. glaberrima*	1	31313547	22303314	TTA	0	0
335	I	*DTT_SAF*	*O. glaberrima*	1	3139572	2414996	TA	0	0
336	I	*DTH_TG*	*O. glaberrima*	1	3178114	2430641	TAG	0	0
337	I	*DTH_TO*	*O. glaberrima*	1	32063578	22843354	TAA	0	0
338	I	*DTH_TO*	*O. glaberrima*	1	3259560	2482799	TTA	0	0
339	I	*DTH_TAA*	*O. glaberrima*	1	33276263	28018512	TAA	0	0
340	I	*DTT_SA*	*O. glaberrima*	1	3339618	2568157	TA	0	0
341	I	*DTT_SAF*	*O. glaberrima*	1	33583786	24151489	TA	0	0
342	I	*DTM_MA*	*O. glaberrima*	1	34065115	24508580	CTGCCTGGCA	0	0
343	I	*DTH_TO*	*O. glaberrima*	1	34822235	25001296	TTA	0	0
344	I	*DTH_TR*	*O. glaberrima*	1	3484655	2718328	TAA	0	0
345	I	*DTH_TO*	*O. glaberrima*	1	3550993	2785464	TAA	0	0
346	I	*DTH_TAJ*	*O. glaberrima*	1	35550255	25669499	TTA	0	0
347	I	*DTH_TAA*	*O. glaberrima*	1	35647397	25747014	ATA	0	0
348	I	*DTT_SC*	*O. glaberrima*	1	3594651	2831363	TA	0	0
349	I	*DTH_TAK*	*O. glaberrima*	1	3706080	22299004	AAT	0	0
350	I	*DTT_SJ*	*O. glaberrima*	1	37296379	27207210	TA	0	0
351	I	*DTH_TR*	*O. glaberrima*	1	3732134	2969271	TTA	0	0
352	I	*DTT_SJ*	*O. glaberrima*	1	3812501	3018181	TA	0	0
353	I	*DTT_SAF*	*O. glaberrima*	1	3814401	3020306	TA	0	0
354	I	*DTH_TC*	*O. glaberrima*	1	3836196	3042879	TTA	0	0
355	I	*DTA_MI*	*O. glaberrima*	1	38506328	28327375	ACTGGGGC	0	0
356	I	*DTH_TO*	*O. glaberrima*	1	38530142	28352552	TAA	0	0
357	I	*DTH_TAL*	*O. glaberrima*	1	3853727	3062262	(C/T)AA	0	0
358	I	*DTT_SJ*	*O. glaberrima*	1	39049480	28844172	TA	0	0
359	I	*DTM_MG*	*O. glaberrima*	1	39500304	29246445	TTGGGTCCC	0	0
360	I	*DTH_TAA*	*O. glaberrima*	1	39810394	29545484	TTA	0	0
361	I	*DTA_HG*	*O. glaberrima*	1	40648068	30313114	GTGCCAAC	0	0
362	I	*DTT_SG*	*O. glaberrima*	1	40744936	30406282	TA	0	0
363	I	*DTT_SD*	*O. glaberrima*	1	4109883	3324756	TA	0	0
364	I	*DTM_MA*	*O. glaberrima*	1	41353649	30791667	GTATATGTA	0	0
365	I	*DTH_TO*	*O. glaberrima*	1	41474060	30909102	TAA	0	0
366	I	*DTT_SAF*	*O. glaberrima*	1	41481931	30917534	TA	0	0
367	I	*DTT_SJ*	*O. glaberrima*	1	42035990	31336175	TA	0	0
368	I	*DTH_TC*	*O. glaberrima*	1	42140161	31410262	TTA	0	0
369	I	*DTH_TAA*	*O. glaberrima*	1	42273321	31537912	TTA	0	0
370	I	*DTH_TW*	*O. glaberrima*	1	42381359	31633231	TTT	0	0
371	I	*DTT_SH*	*O. glaberrima*	1	42694478	31840334	TA	0	0
372	I	*DTT_SW*	*O. glaberrima*	1	42696092	31843109	TA	0	0
373	I	*DTH_TF*	*O. glaberrima*	1	42696204	31843221	TCA	0	0
374	I	*DTH_TG*	*O. glaberrima*	1	43069280	32174013	AGA	0	0
375	I	*DTH_TC*	*O. glaberrima*	1	43468568	32488744	TAA	0	0
376	I	*DTH_TX*	*O. glaberrima*	1	445049	347768	TTA	0	0
377	I	*DTT_SG*	*O. glaberrima*	1	454212	23075768	TA	0	0
378	I	*DTT_SH*	*O. glaberrima*	1	5138352	4145127	TA	0	0
379	I	*DTH_TE*	*O. glaberrima*	1	6436479	4762202	TTA	0	0
380	I	*DTM_MA*	*O. glaberrima*	1	6513527	4837806	CTTATCCAG	0	0
381	I	*DTT_SG*	*O. glaberrima*	1	655715	562531	TA	0	0
382	I	*DTH_TW*	*O. glaberrima*	1	6672110	4982838	TGA	0	0
383	I	*DTT_SH*	*O. glaberrima*	1	6712193	5023409	TA	0	0
384	I	*DTM_MA*	*O. glaberrima*	1	6993028	5274232	ATCATCAGG	0	0
385	I	*DTH_TAI*	*O. glaberrima*	1	7099266	5392844	TAA	0	0
386	I	*DTT_SAF*	*O. glaberrima*	1	7688404	5846504	TA	0	0
387	I	*DTT_SAF*	*O. glaberrima*	1	8199790	6312623	TA	0	0
388	I	*DTH_TO*	*O. glaberrima*	1	8211756	6325553	TTA	0	0
389	I	*DTT_SH*	*O. glaberrima*	1	8315235	6420875	TA	0	0

-

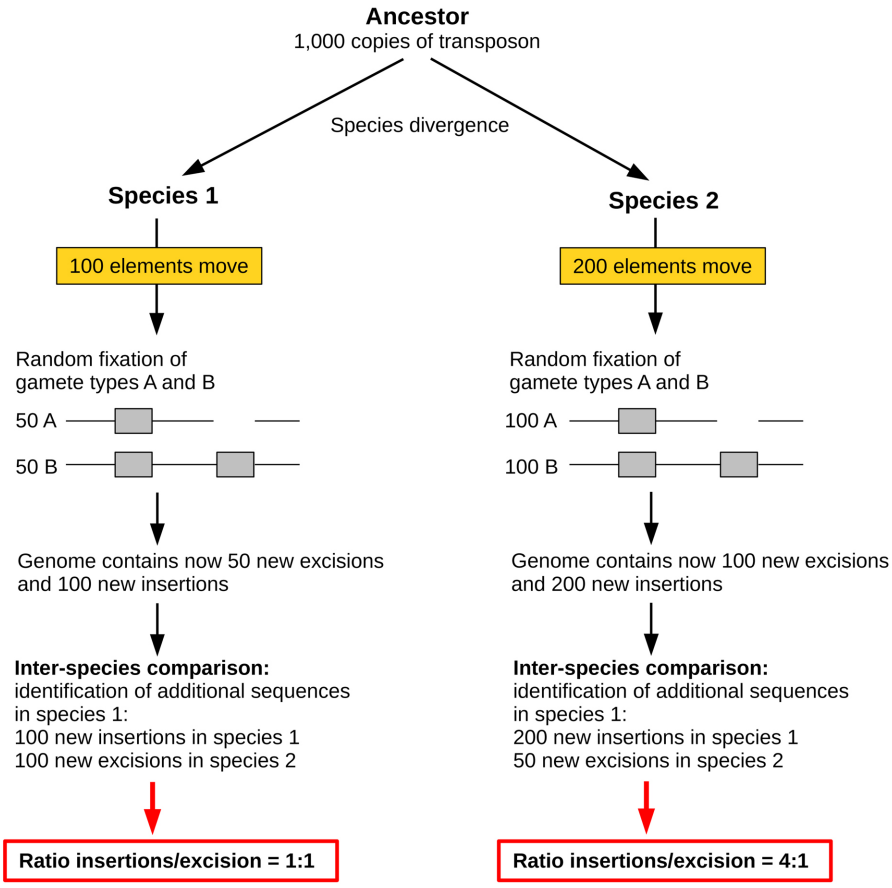
390	I	*DTT_SC*	*O. glaberrima*	1	8405774	6513637	(C/T)A	0	0
391	I	*DTH_TC*	*O. glaberrima*	1	9105923	7143130	GAA	0	0
392	I	*DTT_SS*	*O. glaberrima*	2	10553904	9437687	TA	0	0
393	I	*DTT_SN*	*O. glaberrima*	2	10896325	9736395	TA	0	0
394	I	*DTM_MA*	*O. glaberrima*	2	1097217	938744	AATGCATAA	0	0
395	I	*DTT_SC*	*O. glaberrima*	2	11169256	9976064	TA	0	0
396	I	*DTT_SAF*	*O. glaberrima*	2	1394755	1210004	TA	0	0
397	I	*DTT_SJ*	*O. glaberrima*	2	1441626	1258434	TA	0	0
398	I	*DTH_TAA*	*O. glaberrima*	2	14666538	12263634	TTA	0	0
399	I	*DTT_SC*	*O. glaberrima*	2	15674241	13440855	TA	0	0
400	I	*DTT_SI*	*O. glaberrima*	2	17168914	14466443	TA	0	0
401	I	*DTT_SG*	*O. glaberrima*	2	17192207	14497769	TA	0	0
402	I	*DTT_SAC*	*O. glaberrima*	2	17196700	14503820	TA	0	0
403	I	*DTH_TC*	*O. glaberrima*	2	17225372	14531805	TAA	0	0
404	I	*DTH_TAH*	*O. glaberrima*	2	17330937	14599964	AAT	0	0
405	I	*DTH_TX*	*O. glaberrima*	2	18602093	15133289	TTA	0	0
406	I	*DTH_TG*	*O. glaberrima*	2	18714532	15238083	TTA	0	0
407	I	*DTT_SC*	*O. glaberrima*	2	18743498	15260630	TA	0	0
408	I	*DTH_TAA*	*O. glaberrima*	2	19230476	15622651	T(C/T)A	0	0
409	I	*DTM_MA*	*O. glaberrima*	2	19443668	15840060	TCATAGGAC	0	0
410	I	*DTH_OsKong*	*O. glaberrima*	2	19593695	15944034	(C/T)AA	0	0
411	I	*DTT_SJ*	*O. glaberrima*	2	1970138	1827656	TA	0	0
412	I	*DTH_TO*	*O. glaberrima*	2	19833459	16144180	TTA	0	0
413	I	*DTT_SAF*	*O. glaberrima*	2	19840237	16154011	T(G/A)	0	0
414	I	*DTM_MS*	*O. glaberrima*	2	19966452	16281051	TTTCCTGGG	0	0
415	I	*DTH_TS*	*O. glaberrima*	2	19973007	16291414	TAA	0	0
416	I	*DTH_TR*	*O. glaberrima*	2	20255974	16590745	TTA	0	0
417	I	*DTT_SC*	*O. glaberrima*	2	20890086	17100541	TA	0	0
418	I	*DTH_TW*	*O. glaberrima*	2	22071350	18147549	TAA	0	0
419	I	*DTT_SAF*	*O. glaberrima*	2	22550672	20584672	TA	0	0
420	I	*DTH_TF*	*O. glaberrima*	2	23109331	18782894	TAA	0	0
421	I	*DTH_TA*	*O. glaberrima*	2	23448711	18987739	TAA	0	0
422	I	*DTT_SH*	*O. glaberrima*	2	23463034	19000706	TA	0	0
423	I	*DTT_SA*	*O. glaberrima*	2	23483987	19014058	TA	0	0
424	I	*DTT_SC*	*O. glaberrima*	2	23483987	19014058	TA	0	0
425	I	*DTH_XAB*	*O. glaberrima*	2	23483914	19013985	TTA	0	0
426	I	*DTT_SA*	*O. glaberrima*	2	23875953	19378042	TA	0	0
427	I	*DTT_SJ*	*O. glaberrima*	2	24223170	19693947	TA	0	0
428	I	*DTT_SA*	*O. glaberrima*	2	24556077	19946822	TA	0	0
429	I	*DTT_SC*	*O. glaberrima*	2	24788509	20153594	TA	0	0
430	I	*DTT_SA*	*O. glaberrima*	2	26269326	21160525	TA	0	0
431	I	*DTH_TS*	*O. glaberrima*	2	26298533	21190083	TAA	0	0
432	I	*DTH_TR*	*O. glaberrima*	2	2790606	2537937	TTA	0	0
433	I	*DTH_TC*	*O. glaberrima*	2	28324273	22784202	TCA	0	0
434	I	*DTH_TW*	*O. glaberrima*	2	29853205	23836422	TAA	0	0
435	I	*DTT_SA*	*O. glaberrima*	2	30381782	24334798	TA	0	0
436	I	*DTT_SG*	*O. glaberrima*	2	30595276	24547218	TA	0	0
437	I	*DTT_SQ*	*O. glaberrima*	2	30598443	24548787	TA	0	0
438	I	*DTM_MA*	*O. glaberrima*	2	30852649	24772080	TACTTAAAT	0	0
439	I	*DTH_TC*	*O. glaberrima*	2	30931533	24850090	TAA	0	0
440	I	*DTT_SX*	*O. glaberrima*	2	3316804	2998021	TA	0	0
441	I	*DTH_TO*	*O. glaberrima*	2	33403290	26932183	T(T/A)A	0	0
442	I	*DTT_SX*	*O. glaberrima*	2	34012993	27513144	TA	0	0
443	I	*DTH_TC*	*O. glaberrima*	2	34069760	27566980	TCA	0	0
444	I	*DTH_TG*	*O. glaberrima*	2	34390903	27813086	TT(G/A)	0	0
445	I	*DTH_TS*	*O. glaberrima*	2	34795955	28118299	ATA	0	0
446	I	*DTH_TAA*	*O. glaberrima*	2	35140828	28433640	TAA	0	0
447	I	*DTT_SC*	*O. glaberrima*	2	35499378	28755574	TA	0	0
448	I	*DTH_TC*	*O. glaberrima*	2	35775691	28976707	TCA	0	0
449	I	*DTH_TAA*	*O. glaberrima*	2	3757655	3442028	GAG	0	0
450	I	*DTT_SG*	*O. glaberrima*	2	408889	305338	TA	0	0
451	I	*DTH_TO*	*O. glaberrima*	2	4363382	3914524	TAA	0	0
452	I	*DTH_TO*	*O. glaberrima*	2	4749310	4315470	TAA	0	0
453	I	*DTH_TC*	*O. glaberrima*	2	5310357	4886128	GAA	0	0
454	I	*DTH_TF*	*O. glaberrima*	2	5380801	4954732	ATT	0	0
455	I	*DTH_TAU*	*O. glaberrima*	2	5657125	5242755	TAA	0	0
456	I	*DTT_SD*	*O. glaberrima*	2	5855634	5442020	TA	0	0
457	I	*DTH_TO*	*O. glaberrima*	2	5942793	5538952	TTA	0	0
458	I	*DTT_SI*	*O. glaberrima*	2	5955249	5569352	TA	0	0
459	I	*DTT_SH*	*O. glaberrima*	2	6252462	5839464	TA	0	0
460	I	*DTT_SA*	*O. glaberrima*	2	6314603	5917863	TA	0	0
461	I	*DTM_MA*	*O. glaberrima*	2	6762742	6141194	TACATATGG	0	0
462	I	*DTM_MD*	*O. glaberrima*	2	6783930	6161308	TTAAGGAAA	0	0
463	I	*DTT_SH*	*O. glaberrima*	2	7015546	6403859	TA	0	0
464	I	*DTT_SG*	*O. glaberrima*	2	7026769	6420238	TA	0	0
465	I	*DTH_TW*	*O. glaberrima*	2	7071607	6467288	T(T/A)A	0	0
466	I	*DTT_SC*	*O. glaberrima*	2	7305178	6709042	TA	0	0
467	I	*DTH_TC*	*O. glaberrima*	2	7760792	7165361	TTA	0	0
468	I	*DTM_MT*	*O. glaberrima*	2	8021904	7474342	TACCATTATGTA	0	0
469	I	*DTH_TB*	*O. glaberrima*	2	8299451	7733908	TTA	0	0

-

470	I	*DTT_SA*	*O. glaberrima*	2	8317260	7752000	TA	0	0
471	I	*DTM_MA*	*O. glaberrima*	2	9070596	8405297	TATTTATAA	0	0
472	I	*DTT_SE*	*O. glaberrima*	2	9118930	8435181	TA	0	0
473	I	*DTT_SH*	*O. glaberrima*	2	9411322	8553775	TA	0	0
474	I	*DTT_SA*	*O. glaberrima*	2	943835	801480	TA	0	0
475	I	*DTH_TAA*	*O. glaberrima*	3	10434766	9645591	(C/T)GA	0	0
476	I	*DTH_TW*	*O. glaberrima*	3	10925359	10103886	TCA	0	0
477	I	*DTH_TG*	*O. glaberrima*	3	10959877	10138641	ATA	0	0
478	I	*DTH_TS*	*O. glaberrima*	3	11460996	10512412	TTA	0	0
479	I	*DTH_TO*	*O. glaberrima*	3	11563899	10631536	TTA	0	0
480	I	*DTT_SI*	*O. glaberrima*	3	12372363	11394352	TA	0	0
481	I	*DTT_SG*	*O. glaberrima*	3	12372711	11394939	TA	0	0
482	I	*DTH_TAD*	*O. glaberrima*	3	12397058	11419392	TAG	0	0
483	I	*DTM_MAC*	*O. glaberrima*	3	12399247	11421625	TTTTTTTAA	0	0
484	I	*DTT_SG*	*O. glaberrima*	3	12435261	11463175	TA	0	0
485	I	*DTH_TAB*	*O. glaberrima*	3	12915389	11939220	TTA	0	0
486	I	*DTH_TW*	*O. glaberrima*	3	12931333	11956232	TAC	0	0
487	I	*DTM_HA*	*O. glaberrima*	3	12940204	11967304	CACCGAGAC	9	0



Supplementary Figure S3. Inheritance of transposon insertion/excision patterns. For this model we assume that all transposition effects are selectively neutral. It is commonly accepted that on mechanism of multiplication is for DNA transposons to excise during DNA replication and to re-insert in front of the replication fork. This leads to one daughter strand with one copy of the element (A-type gamete) and one with two copies (B-type gamete). If a large number of transposons are active in many different loci in a species (this may be spread out over many generations), the offspring genome will be a mosaic of loci derived from A and B-type gametes. When comparing that genome to that of a closely related species, loci resulting from A-type gametes will identify an excision and an insertion, while loci resulting from A-type gametes will only identify insertions. Thus the observed overall ratio of insertions to excisions fro a given transposon family will be 2:1.



Supplementary Figure S4. Detection of differences in transposon activity in different species. This model assumes that a given transposon family was present in many copies in the ancestor species. After species divergence, the transposon family is active at different levels in the two species (100 transpositions in one and 200 in the other species). As described in Supplementary Figure 1, A and B-type gametes are passed on to offspring in a 1:1 ratio. In a cross-species comparison which identifies transposons (additional sequences) which are present in one but absent in the other species, insertions in one species and excisions in the other will be detected. If a transposon family had different levels of activity in the two species since their divergence, insertion/excision ratios will deviate from the 2:1 ratio.

Chapter 4:

The making of a genomic parasite – the *Mothra* family sheds light on the evolution of *Helitrons* in plants

The following chapter describes the *DHH_Mothra* family, a ubiquitous *Helitron* in rice. We describe the possible evolution from an autonomous to non-autonomous element, involving several steps. Moreover, we could demonstrate that the RPA homolog of plant *Helitrons* was most likely acquired by horizontal transfer. This work was published by Roffler *et al.* in *Mobile DNA 2015.*

Roffler et al. Mobile DNA (2015) 6:23
DOI 10.1186/s13100-015-0054-4

Mobile DNA

RESEARCH Open Access

The making of a genomic parasite - the *Mothra* family sheds light on the evolution of *Helitrons* in plants

Stefan Roffler, Fabrizio Menardo and Thomas Wicker*

Abstract

Background: Helitrons are Class II transposons which are highly abundant in almost all eukaryotes. However, most Helitrons lack protein coding sequence. These non-autonomous elements are thought to hijack recombinase/helicase (RepHel) and possibly further enzymes from related, autonomous elements. Interestingly, many plant Helitrons contain an additional gene encoding a single-strand binding protein homologous to Replication Factor A (RPA), a highly conserved, single-copy gene found in all eukaryotes.

Results: Here, we describe the analysis of *DHH_Mothra*, a high-copy non-autonomous Helitron in the genome of rice (*Oryza sativa*). *Mothra* has a low GC-content and consists of two distinct blocs of tandem repeats. Based on homology between their termini, we identified a putative mother element which encodes an RPA-like protein but has no *RepHel* gene. Additionally, we found a putative autonomous sister-family with strong homology to the *Mothra* mother element in the RPA protein and terminal sequences, which we propose provides the RepHel domain for the *Mothra* family. Furthermore, we phylogenetically analyzed the evolutionary history of RPA-like proteins. Interestingly, plant Helitron RPAs (PHRPAs) are only found in monocotyledonous and dicotyledonous plants and they form a monophyletic group which branched off before the eukaryotic "core" RPAs.

Conclusions: Our data show how erosion of autonomous Helitrons can lead to different "levels" of autonomy within Helitron families and can create highly successful subfamilies of non-autonomous elements. Most importantly, our phylogenetic analysis showed that the PHRPA gene was most likely acquired via horizontal gene transfer from an unknown eukaryotic donor at least 145–300 million years ago in the common ancestor of monocotyledonous and dicotyledonous plants. This might have led to the evolution of a separate branch of the Helitron superfamily in plants.

Keywords: Transposon, Helitron, RPA, Rice, Horizontal transfer

Background

Helitrons are a superfamily of transposable elements (TEs) in eukaryotes which was discovered only relatively recently in *Arabidopsis thaliana*, *Caenorhabditis elegans* and *Oryza sativa* [1]. They have since been found in many genomes of flowering plants [1, 2], mosses [3], fungi [4–6] but also many animals such as sea urchin [7], fish [8, 9] and bats [10]. A recent *in silico* analysis using the program *Helsearch* [2] estimates the number of Helitrons in rice and sorghum to approximately 7000 and 5000, respectively, covering several megabases of their hosts' genomes. The most extensively studied genome regarding Helitrons is the one of maize, where approximately 2000 intact Helitrons and more than 20,000 Helitron fragments were found. Based on high homology between individual elements they are thought to still be very active [11]. As for most DNA transposons, the majority of Helitron elements are non-autonomous and do not encode any proteins. These non-autonomous elements presumably depend for their transposition on enzymes encoded by "mother" or "master" elements elsewhere in the genome.

One reason why Helitrons remained undiscovered for a long time is their limited diagnostic features. They lack

* Correspondence: wicker@botinst.uzh.ch
Institute of Plant Biology, University of Zürich, Zollikerstrasse 107, Zürich CH-8008, Switzerland

terminal inverted repeats (TIRs) and the only motifs common to all Helitrons are the dinucleotide TC at the 5' end as well as a CTRR motif at the 3' end. Additionally, almost all Helitrons have a G/C rich 15–20 bp hairpin motif approximately 10–12 bp upstream of the 3' end, which is thought to serve as a stop signal in the transposition process [1]. Finally, Helitrons have a strong preference to insert between the bases A and T or sometimes between two Ts [1].

The transposition mechanism of Helitrons and the involved proteins differ from those of the well described DDE transposases. Autonomous Helitrons encode a RepHel protein of 1000–3000 amino acids (aa) length, which is thought to initiate the replication. The RepHel constitutes a replication initiation domain (RCR/Rep) followed by a helicase enzyme (Hel) of approximately 400 aa [12]. Because of structural homology with the catalytic core of HUH endonucleases of a bacterial rolling-circle transposons [13], it was suggested that Helitrons use a rolling-circle mechanism involving a single-stranded DNA intermediate for transposition and replication [1, 12]. Li and Dooner [14], however, clearly showed excisions of Helitrons from 0.4 to 6 kb size in somatic Maize tissue. This challenges the current model and suggests an alternative mode of transposition involving excision and repair similar to TIR transposons. Indeed, it is possible that single stranded DNA transposition can result in the elimination of that copy from that locus when occurring during S phase of meiosis 1 [15].

Even though Helitrons are ascribed to the Class II (DNA) transposons, they remain unique due to their exclusive structural features and transposition mechanism and belong to a separate subclass within the DNA transposons [16]. However, rolling-circle transposition mechanisms have been described for gemini viruses [17], plasmids and some bacterial transposons [18]. Structural homology between their transposases suggests very ancient origin of Helitrons [1].

In plants, some Helitrons have been reported to also encode a distant homolog of the Replication Protein A (RPA), a protein ubiquitous in eukaryotes [19, 20]. RPA has several single-strand DNA binding sites and is involved in processes such as DNA replication and repair. RPA homologs have also been identified in Helitrons from zebrafish and sea anemone [12] and in Helentrons (a sub-type of Helitrons) in *Drosophila melanogaster* [21].

At least in maize, Helitrons seem to acquire close by gene fragments very frequently. Several studies showed an ongoing gene movement, gene shuffling and transcriptional read-troughts, which is attributed to Helitron activity [22, 23]. In the maize line B73, approximately 11,000 such chimeric transcripts have been found to be expressed which represents almost one quarter of all genes [24]. Therefore, it is thought that Helitrons contributed substantially to the recent diversification observed in the maize genus. Moreover, frequent gene capturing mediated by Helitrons was also reported in the silk worm *Bombyx mori* [25] and in the bat *Myotis lucifugus* [26].

In this study we describe the analysis and origin of a high-copy Helitron family in rice, which we named *DHH_Mothra.* Non-autonomous *Mothra* elements are present in hundreds or even thousands of copies in multiple rice species, which merited an in-depth analysis of this TE family. We identified a putative mother element for the *Mothra* family that encodes an RPA homolog but no RepHel protein. We moreover identified a closely related Helitron family, which we propose to be the donor for the lacking RepHel enzyme of *Mothras.* According to our model, this introduces an additional level of autonomy. We furthermore investigated the evolutionary background of Helitron RPA acquisition in plants and suggest horizontal transfer most likely from a unicellular eukaryote into the common ancestor of mono- and dicotyledonous plants.

Results

Mothra is a high-copy non-autonomous Helitron

In a previous study [27] we compared the two closely related rice species *O. sativa*, the Asian rice, with its relative *O. glaberrima*, the African rice, and investigated presence/absence polymorphisms of Class II transposons of the TIR subclass. While scanning polymorphic TE sites, we repeatedly encountered a sequence which was obviously of repetitive nature but we were unable to classify it at that time. Now, we found that it was in fact a non-autonomous TE of the Helitron order which we called *Mothra*.

We identified a total of 1,682 *Mothra* elements from which we manually deduced consensus sequences of 22 sub-types. The 22 *Mothra* sub-types share the same terminal and internal sequence motifs but vary in size between 1252 and 2741 bp (see Methods). The differences in size between the sub-types are due to differences in the order, length and/or orientation of blocs of tandem repeats (see below). From these 22 sub-types, we created a single consensus sequence of 1993 bp in length which we refer to as consensus of the non-autonomous *Mothra* elements (Fig. 1a). As described for other Helitrons, *Mothra* elements show the characteristic dinucleotide TC at its 5' end and the four bases CTAG at the 3' end. Additionally, we found the characteristic hairpin motif of 16 bp length located 13 bp upstream of the 3' end of the elements. From this, we concluded that *Mothra* is in fact is a non-autonomous TE of the Helitron order.

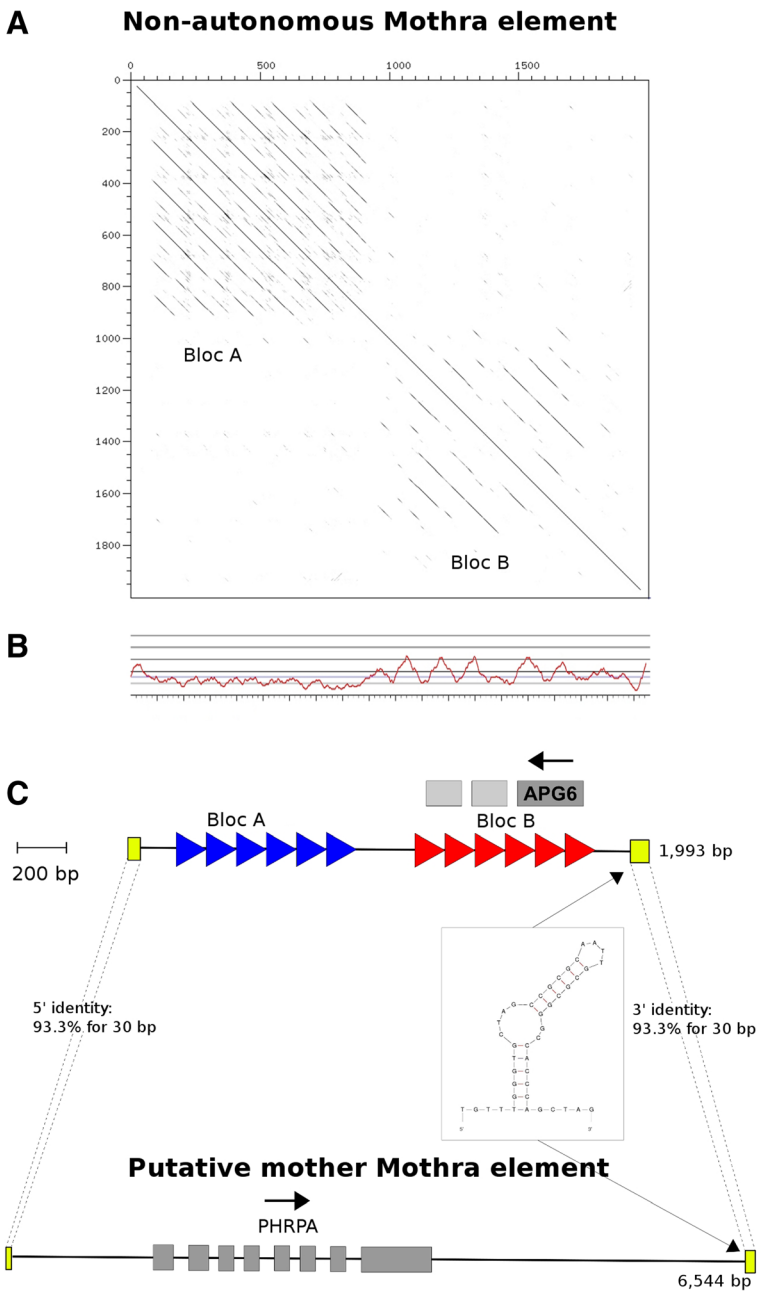


Fig. 1 Overview of the non-autonomous *Mothra* consensus sequence and its putative mother element. **a** Dot-plot of the non-autonomous *Mothra* consensus sequence against itself shows the two repetitive Blocs A and B. **b** GC-plot of the non-autonomous element. Note that Bloc A shows a unusual low GC-content of approximately 20 %. **c** Schematic overview of the non-autonomous *Mothra* and its putative mother element below. Both elements share the characteristic hairpin structure at the 3' end. The termini of the putative mother element and the non-autonomous consensus are conserved (in yellow). Furthermore, the non-autonomous elements shows the putative ORF of 96 amino acids. Note here, that the putative mother element of *Mothras* encodes for a RPA homolog, which we named PHRPA, but no RepHel protein

Mothra contains tandem repeats and gene fragments

Mothra contains two distinct sequence blocs (Bloc A and B, Fig. 1a). Bloc A, which ranges approximately from position 80 to position 900 in the consensus sequence, consists of six direct repeats and shows a very low GC content of 20 %. Bloc B ranges from position 950 – 1860 and consists of six different, less conserved direct repeats and exhibits an average GC content of about 40 % (Fig. 1b). There is great variety in the number of the repeat units within the Blocs A and B among the

individual copies. In some cases, the order of the blocs is even reversed. In other cases, additional sequence is present between or sometimes even within one of the two blocs.

By definition, non-autonomous elements do not encode any proteins. But interestingly, the *Mothra* consensus sequence contains a putative open reading frame (ORF) of 96 amino acids in reverse orientation in Bloc B. The predicted protein shows sequence homology to the APG6 domain (Pfam ID: pfam04111, e-value: 2,2e-03) which has been described to be involved in autophagy and vascular sorting pathways in yeast [28]. Because of the repeat structure of Bloc B, this homology is partially repeated two more times downstream of this ORF. These additional copies, however, lack start codons and therefore do not constitute intact ORFs. We assume that this ORF is the result of gene fragment capture but probably has no function. The fact that this gene fragment is part of the *Mothra* consensus sequence indicates that the gene capture event occurred before the radiation of the *Mothra* family.

The putative *Mothra* mother element lacks a *RepHel* gene

Usually, non-autonomous TEs share their terminal sequences with their autonomous "mother" elements. That is why we scanned the genome of *O. sativa* using the first 50 and the last 80 bp of the non-autonomous element, respectively, as queries. We extracted 323 sequences where we identified both ends in the same orientation located within 25 kb from each other. We scanned the 323 fragments for the presence of transposases and helicases but could not identify a single one. However, we identified one sequence of 6544 bp in length that encodes an RPA homolog (Fig. 1c). This RPA sequence was annotated in the rice genome as hypothetical protein (LOC_Os11g47400). The predicted protein contains several generic single-stranded DNA-binding sites. After manual re-annotation of the protein we were able to extend the putative protein length from 296 aa to 472 aa and the number of exons from four to eight. Interestingly, this sequence was the only one among all 323 analyzed fragments containing a putative complete gene between the two *Mothra* ends. The sequence homology between the termini of this putative mother element and the non-autonomous *Mothra* consensus is very high (93,3 % of the terminal 30 bp, and 81,2 % and 80,2 %, respectively for the terminal 100 bp). According to Yang et al. [2], this makes them not only members of the same family but also of the same sub-family. Moreover, we identified a deletion derivative of the putative mother element that shows homology to almost the entire element but lacks the RPA domain (Fig. 2). This indicates that we indeed identified a distinct element rather than an RPA homolog that is flanked by chance by two fragments of termini from non-autonomous *Mothra* elements. Therefore, we propose this element, even if we did not find an ORF encoding an RepHel protein, to be the mother element of the numerous non-autonomous *Mothras*. Thus, in the strict sense, the putative *Mothra* mother element might itself not be autonomous (see before).

Polymorphisms between *O. sativa* and *O. glaberrima* demonstrate recent activity of *Mothra* elements

In a previous study we produced an alignment of approximately 60 % of the *O. sativa* and *O. glaberrima* genomes for identification of presence/absence polymorphisms of TIR transposons [27]. Now, we searched this alignment for polymorphisms related to *Mothra* elements. Out of a total of 856 *Mothra*-related polymorphisms, we investigated 148 manually. Most of them turned out not to be actual presence/absence polymorphisms, but rather variations in the number of repeat units between orthologous *Mothras* of the two species. Most of these differences probably arose from mechanisms such as unequal crossing-over or repeat slippage rather than from transposition activity (Fig. 3). Thus, the vast majority of *Mothra* copies are found in the same position in both rice species,

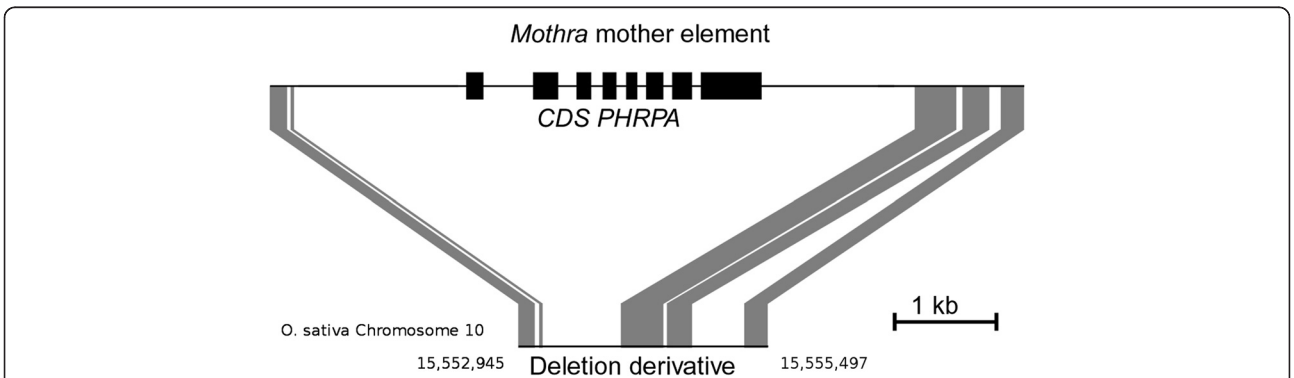


Fig. 2 Schematic representation of an identified deletion derivative of the putative mother element of *Mothras*. The coding sequence for the *PHRPA* gene is indicated in black while the homologous sequences are indicated in gray

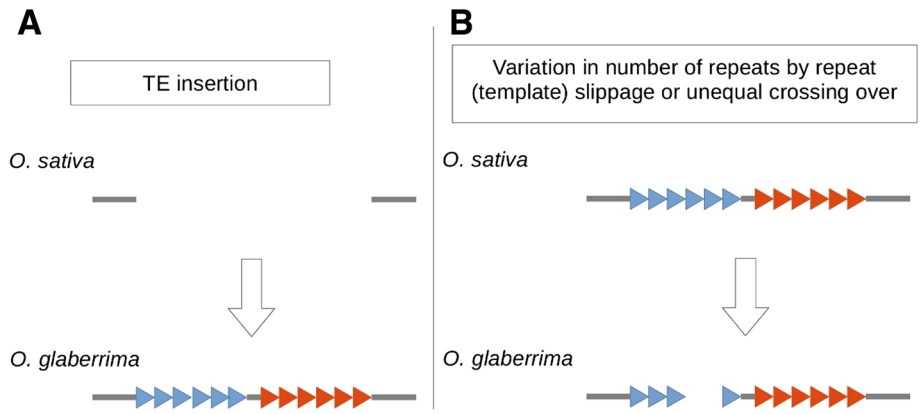


Fig. 3 Presence/absence polymorphisms of *Mothra* elements between *O. sativa* and *O. glaberrima*. *O. sativa* and *O. glaberrima* diverged approximately 600.000 ya. **a** True polymorphisms are characterized by presence/absence of whole elements. We found a total of four insertions and four putative excisions. **b** Most of the variation in the two *Mothra* populations comes from repeat-slippage or unequal crossing-over

meaning that they inserted before the two species diverged approximately 600,000 years ago [29]. Therefore, we can say that most of the copies are older than 600,000 years.

However, we also identified eight sites where we found putative insertion/excision polymorphisms of non-autonomous *Mothras* between the two rice species (Fig. 4a). In four cases, we found the *Mothra* element located between the characteristic nucleotides A and T present in *O. sativa* but not in *O. glaberrima.* Because Helitrons do not generate target site duplications, these events probably represent typical insertions in *O. sativa.* Interestingly, we found four sites where we suspect putative *Mothra* excisions. We conclude this based on the DNA repair patterns which are similar to those described for TIR DNA transposon excisions [30] (Fig. 4b). In two cases, we observed incomplete excision events whereas the other two cases went along with a deletion and the introduction of filler DNA, respectively.

The eight polymorphic elements correspond to 5,4 % of subset of 148 manually investigated polymorphisms. Considering that we identified a total of 856 insertion/deletion polymorphisms between the two species, we extrapolate that a total of approximately 46 *Mothra* elements have moved since the two species diverged about 600,000 years ago [29]. However, this number is based on approximately 60 % of the genome which was aligned. Thus, the actual number of transposed elements might be even higher. Compared to the previously investigated TIR transposons [27], we conclude that *Mothra* has a level of activity similar to that of highly active DTT-Mariner elements.

Phylogenetic analysis of the *Mothra* RPA homolog family

RPA proteins are involved in crucial processes such as DNA-replication and -repair. Furthermore, this "core" RPA is a single copy gene and highly conserved among eukaryotes. This makes RPA useful for phylogenetic analysis and, thus, to study the origin of the plant Helitron RPA homolog (PHRPA). We used the the original "core" RPA as well as identified *Mothra* PHRPA of *O. sativa* as queries for NCBI blast searches against representatives from all major eukaryotic branches. We also included species from the largely under-sampled unicellular eukaryotic clades, such as Alveolata, Amoebae, Oomycetes and Rhizaria. Furthermore, we include two RPA homologs from Helentrons that were identified in *Drosophila melanogaster* [21] to investigate their relationship to PHRPAs. As an outgroup, we used some distant homologs from archaea (Fig. 5). Except in monocotyledonous and dicotyledonous plants, we usually found exactly one *RPA* gene (see below). The final dataset comprised 72 proteins from 62 species.

Our results show that most major eukaryotic clades cluster in monophyletic groups. We observe a clear grouping into plants, animals, fungi and Oomycetes. The phylogeny within these clades is consistent with the established taxonomy of eukaryotes [31]. For example plant RPAs first split into algae, mosses and later into monocots and dicots (Fig. 5). Because of the robustness of the tree and the great concordance with the taxonomy, these proteins most probably represent the intrinsic, eukaryotic "core" RPAs.

Most clades have exactly one RPA gene but there are exceptions. Interestingly, one of the two copies obtained from the Alveolata, *Cryptosporidium,* also clusters at the root of the plant branch. However, the other copy we find, as expected, in the clade of Alveolates, which are even more distant to the core RPA clade than the PHRPA family. Furthermore, we found two RPA paralogs in the genomes of *Physcomitrella*, a Moss, and the

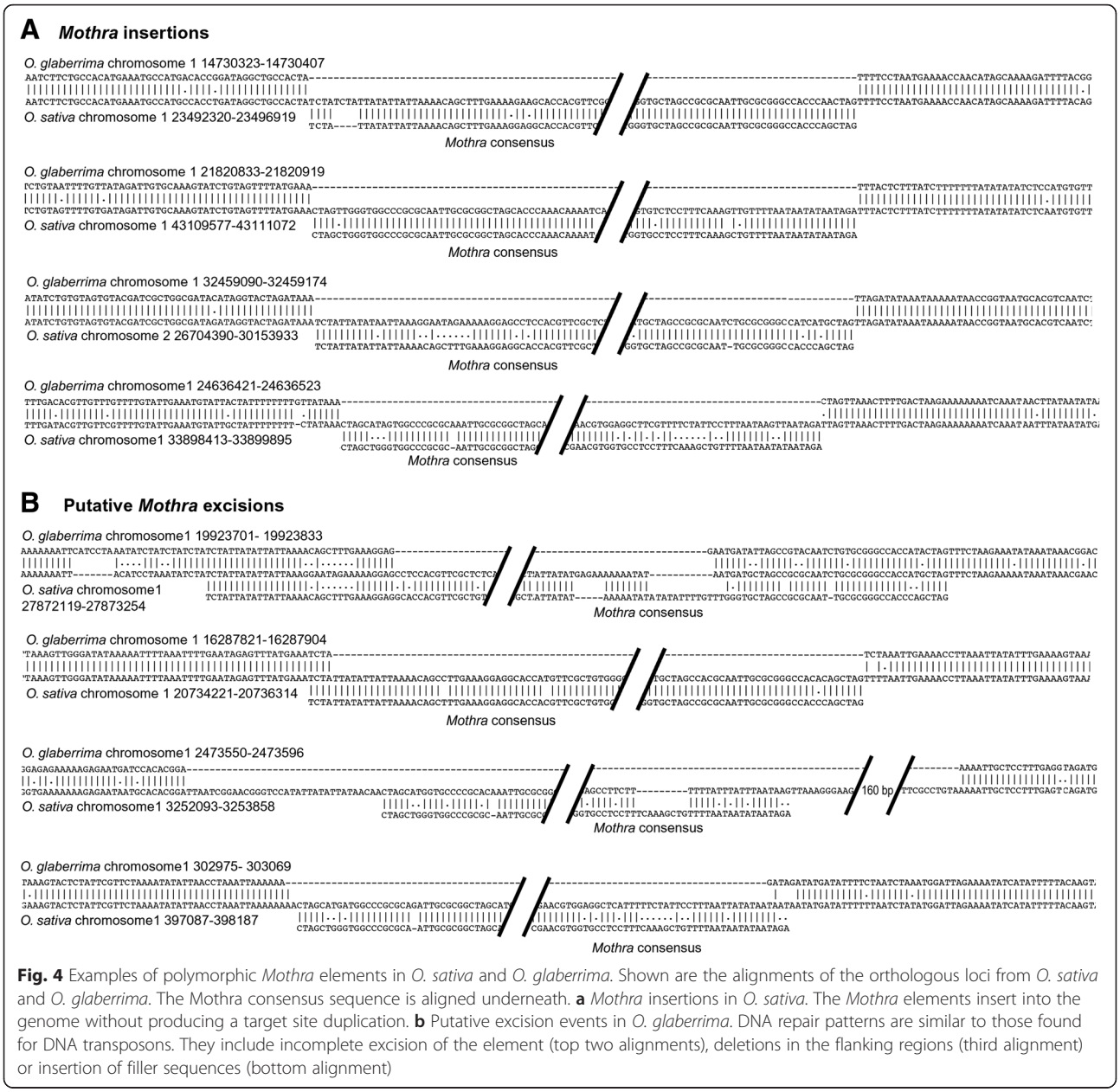


Fig. 4 Examples of polymorphic *Mothra* elements in *O. sativa* and *O. glaberrima*. Shown are the alignments of the orthologous loci from *O. sativa* and *O. glaberrima*. The Mothra consensus sequence is aligned underneath. **a** *Mothra* insertions in *O. sativa*. The *Mothra* elements insert into the genome without producing a target site duplication. **b** Putative excision events in *O. glaberrima*. DNA repair patterns are similar to those found for DNA transposons. They include incomplete excision of the element (top two alignments), deletions in the flanking regions (third alignment) or insertion of filler sequences (bottom alignment)

green alga *Chlorella*, to form a monophyletic group on the same level as the PHRPAs. It is possible that there are contaminations since these organisms are difficult to isolate and cultivate.

Most importantly, we find the PHRPAs to form a separate, monophyletic group outside the core RPA clade. Thus, we conclude that the PHRPA ancestor protein has evolved very early in the transition from prokaryotes to eukaryotes. Interestingly, we only find representatives of mono- and dicotyledonous plants in the PHRPA clade. Moreover, PHRPAs are more diverse than core RPAs. Indeed, PHRPA proteins are on average 21 % identical to each other, while core RPAs show an average of 39 % sequence identity (Additional file 1: Figure S1). Also the branch lengths of the PHRPA clade are noticeable long. This suggests diversification of new, independent gene subfamilies. The possible reasons why these proteins are only found in monocots and dicots which diverged approximately 145–300 million years ago (mya) [32, 33] are discussed below (see discussion). Moreover, our analysis reveals that the RPA proteins acquired by Helentrons seem to be of another, even more distant origin. These proteins form a separate group which branches off before the radiation of eukaryotes (Fig. 5).

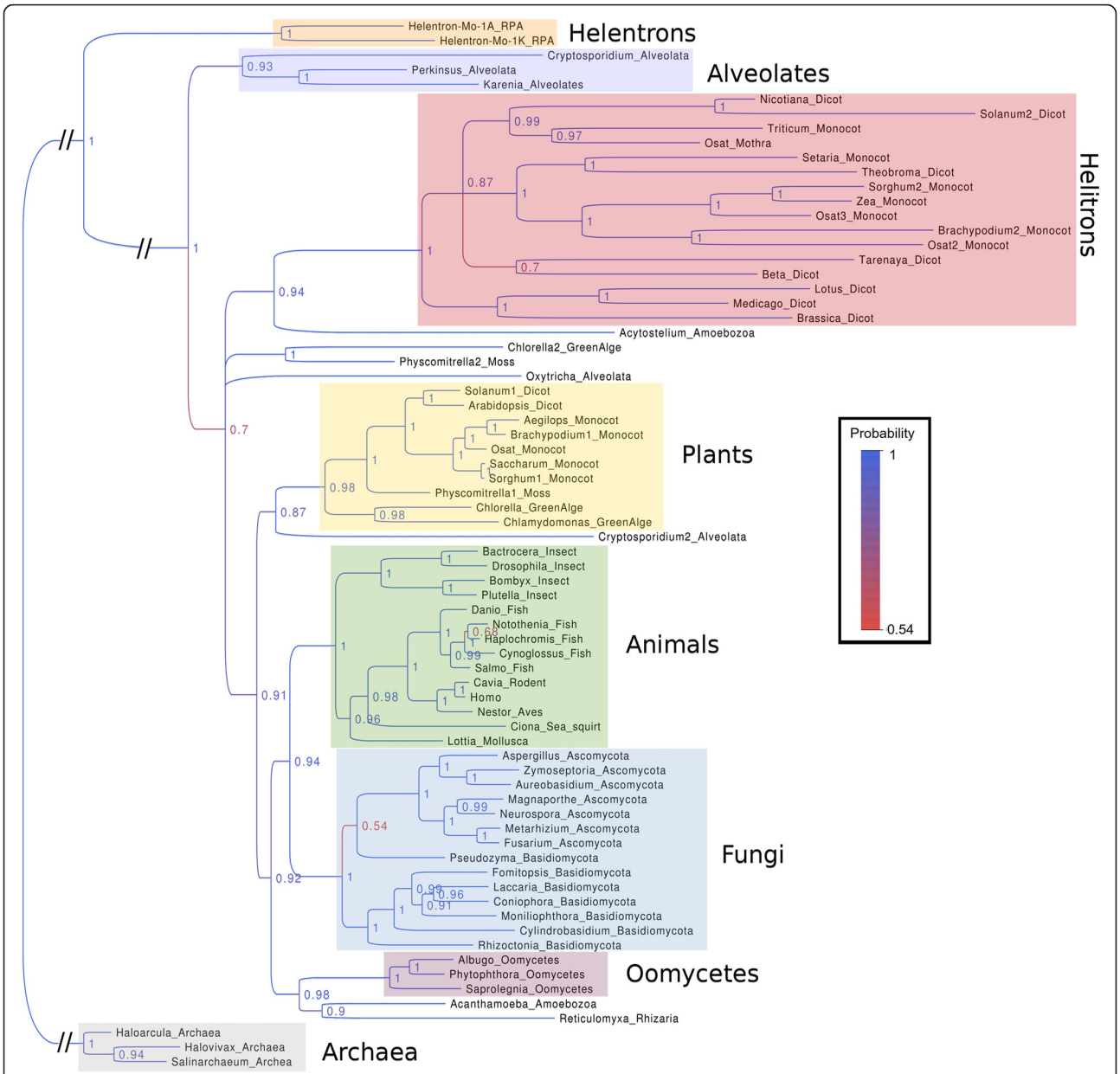


Fig. 5 Phylogenetic tree of 72 RPA homologs from 62 species. We found clear monophyletic grouping of the original RPA protein into the "core" eukaryotic clades such as plants, animals, fungi, Oomycetes and Alveolates. Within the clades the taxonomy widely reflects the diversification within these clades. The Helitron clade, which includes the *Mothra* RPA protein, however, forms an independent branch. Interestingly, it includes only representatives of the monocotyledonous and dicotyledonous plants. For the phylogeny we used a maximum likelihood approach using the software MrBayes (reversible jumping Markov chain Monte Carlo (MCMC) simulations for 4 million generations). The numbers at branches reflect confidence values (i.e. probabilities that sequences to the right of the fork group together in 6000 generated trees)

Mothras might use the RepHel protein of closely related Helitrons

Above, we describe that the putative mother element of the non-autonomous *Mothras* encoded an PHRPA protein but not for a RepHel protein. This raises the question of how these elements would actually transpose. As it has been described for non-autonomous elements, that they recruit closely related transposases, we suspect that RepHel from a closely related Helitron family would be used by *Mothra* elements. Therefore, we scanned the *O. sativa* genome for homologs of the PHRPA protein and extracted 21 fragments including 20 kb up- and downstream of the protein. Out of these we identified nine sequences with sizes from 8064 to 15,513 bp that all contain a *PHRPA* homolog and an adjacent *RepHel* gene.

Based on sequence homology we could clearly differentiate them into three groups. While we found five copies of group 1 elements, there were two copies each for groups 2 and 3, respectively. The PHRPA of group 1 is most similar to that of the Mothra mother element (46.1 % similarity compared to 21.6 % and 22.6 % for groups 2 and 3, respectively). Moreover, the elements of group 1 and *Mothras* nearly fulfill the criteria of Yang et al. [2] to belong to the same family (73 % identity over 30 bp at the 5' end and 77 % identity at the 3' end). Because of this and the strong homology of their RPA proteins, we henceforth refer to these Helitrons of as the sister-family of *Mothra* (Fig. 6). Interestingly, when we compared the five copies of the sister-family with those in *O. glaberrima*, we found all of them to be polymorphic (Table 1), indicating recent activity of the *Mothra* sister-family. Thus, we propose that *Mothra* elements recruit the RepHel protein of their sister-family to transpose. For both, the *PHRPA* gene of the *Mothra* mother element and *PHRPA* and *RepHel* of the sister-family, we found transcripts in NCBI, suggesting that both might still be active (Additional file 2: Table S1).

Discussion

The goal of our study was to characterize the origin and evolution of the high-copy Helitron family *Mothra* in rice. Although Helitrons are found in nearly all eukaryotic genomes they are much less well understood than other TE superfamilies. Despite their considerable role in exon shuffling and gene movement in plants [22–24], only few studies are available that shed light on their transposition mechanism. Initially, it was proposed that Helitrons replicate via a rolling-circle mechanism [1]. However, this was challenged by the discovery of Helitron excisions in somatic maize tissue [20]. Our data also suggest that some of the presence/absence polymorphism in rice might represent Helitron excisions. While Li and Dooner [14] mainly found repair patterns introducing TA micro-satellites as “filler” DNA, our putative excision events were also associated with deletions of the flanking sequences. These footprints strongly resemble those of TIR transposon excisions [27, 29, 34, 35]. Thus, these combined findings suggest the existence of at least one alternative transposition pathway to the proposed rolling-circle mechanism.

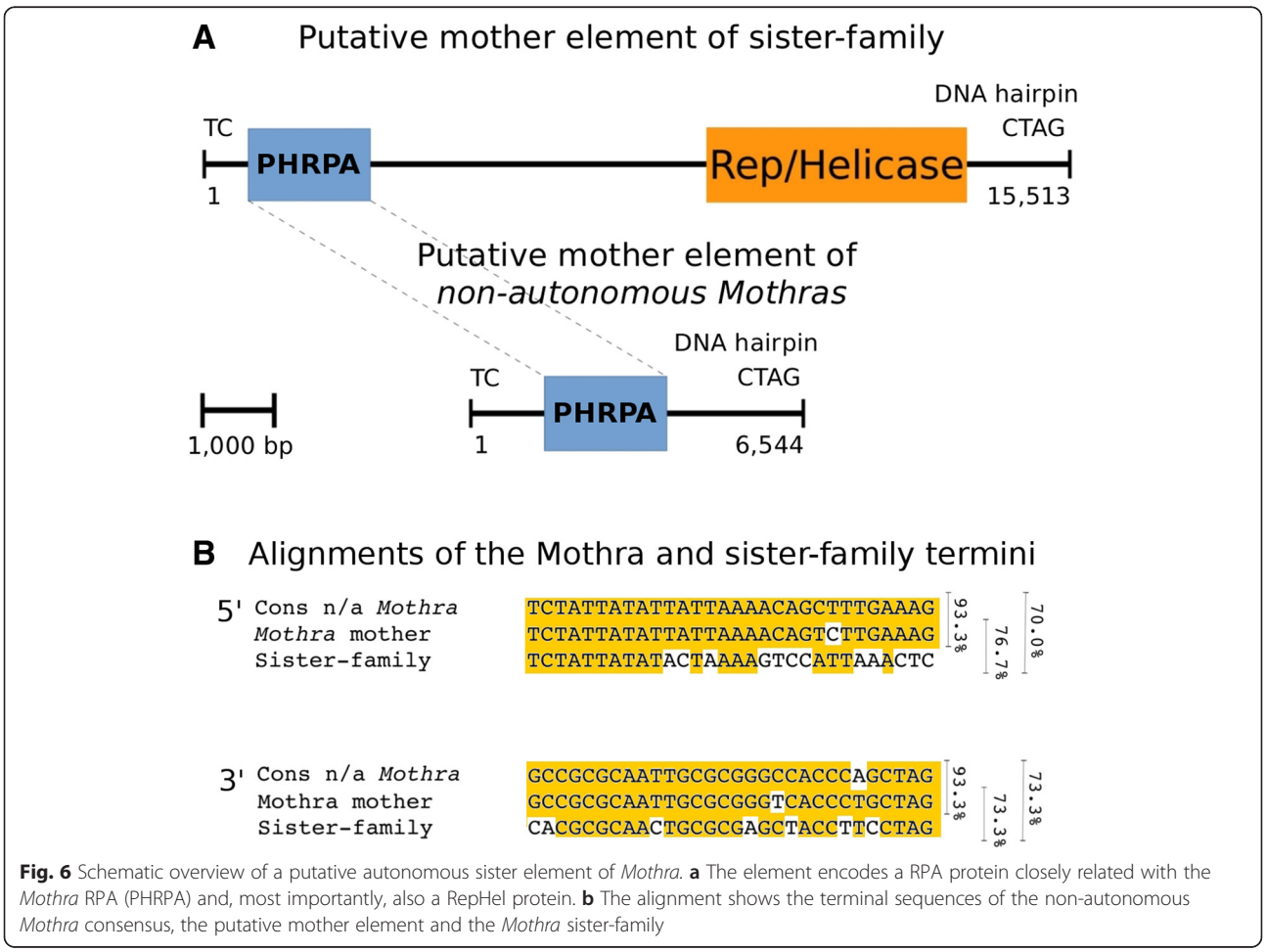


Fig. 6 Schematic overview of a putative autonomous sister element of *Mothra*. **a** The element encodes a RPA protein closely related with the *Mothra* RPA (PHRPA) and, most importantly, also a RepHel protein. **b** The alignment shows the terminal sequences of the non-autonomous *Mothra* consensus, the putative mother element and the *Mothra* sister-family

Table 1 Overview of all identified copies of the putatively autonomous elements of the *Mothra* sister-family in *O. sativa* and *O. glaberrima*

Mothra sister-family copies			
O. sativa			
	Start pos.	End pos.	Comment
Chromosome			
11	26,634,911	26,619,399	Reverse
11	22,184,151	22,168,642	Reverse
11	24,183,680	24,199,188	Forward
5	592,132	606,965	Forward
5	25,964,570	25,949,203	Reverse
O. glaberrima			
	Start pos.	End pos.	Comment
Chromosome			
11	19,460,159	19,467,758	No RPA

Despite these open questions, the main findings of our study provided insight into the evolution of different levels of non-autonomous elements and, more importantly, of the Helitron superfamily in plants in general. Our main conclusions are discussed in the following.

Sequence composition of non-autonomous *Mothras* elements might play a role in transposition efficiency

Non-autonomous transposons can create hundreds or even thousands of copies in only few generations [36]. Loss of protein coding sequences and thereby autonomy has happened in all major Class II TE superfamilies. It can be explained by the fact that hosts regulate TEs via epigenetic silencing. Thus, constant reshaping, shortening and the accumulation of "nonsense" sequences might be mechanisms to avoid RNA silencing [37]. Alternatively, the presence of an active functional copy might release selection pressure on other copies, allowing for non-autonomous derivatives to emerge. Still, non-autonomous elements retain the ability to cross-mobilize related transposases. This type of trans-acting system has best been described in detail for the TIR transposons of the *DTT-Mariner* superfamily [36]. Transient expression experiments in yeast showed that the affinity for the autonomous element was determined by the TIR region. The efficiency of transposition, however, was influenced dramatically, positively or negatively, by different compositions of internal sequences.

We suspect that the great success of *Mothra* elements might have to do with their unusual sequence composition (see Fig. 1). The Blocs A and B of the non-autonomous element are unique to *Mothra* elements and their high conservation within the *Mothra* family suggests functional importance. When we screened the genomes of *O. sativa* and *O. glaberrima* for *Mothra* related polymorphisms (see above), we found that the majority of the differences were variations in the number of repeat units. Most likely these were caused by repeat slippage or unequal crossing-over for which the repeat arrays of Blocs A and B served as templates. Thus, these repeat arrays may be a sources of plasticity and permanent turnover within non-autonomous *Mothra* elements.

The *Mothra* RPA homolog likely originated from horizontal transfer

In our phylogenetic analysis of RPA proteins we found clear monophyletic clustering of the "core" RPAs in all major eukaryotic groups which broadly reflects the separation of early eukaryotes into distinct lineages (see Fig. 5). Interestingly, the clade representing the RPA homologs from plant Helitrons (PHRPAs) branches off even before the separation of plants, animals, fungi and Oomycetes, indicating a very ancient origin of these proteins. It is the more surprising that this clade only includes proteins from monocotyledonous and dicotyledonous plants which only separated approximately 145–300 mya [32, 33]. Previous studies proposed that plant Helitrons hijacked and modified the eukaryotic core *RPA* gene which later became the plant Helitron RPA [1, 38]. However, the clear monophyletic origin of PHRPAs outside the core RPA clade challenges this model.

There are two possible explanations for the phylogenetic position of PHRPAs: First, PHRPA proteins were originally present in all other eukaryotes and were lost in all lineages except the monocots and dicots. We consider this highly unlikely. The second explanation (which we clearly favor) is horizontal gene transfer. Typical characteristics of horizontal gene transfer are phylogenetic incongruence and/or unusually high sequence identity of proteins from otherwise distantly related species. In our case, we found very well supported phylogenetic incongruence. However, we could not identify a putative donor of PHRPA. This donor was obviously not sampled in our collection. We propose that PHRPA was transfered from this unknown and distantly related eukaryote into the progenitor of monocots and dicots. This horizontal transfer must have occurred before monocots and dicots diverged 145–300 mya [32, 33], since we have not found PHRPAs in any other plant group that diverged earlier. Our data indicate that the progenitor of all eukaryotic RPA genes was already present during eukaryogenesis, but it remains unclear if the last eukaryotic common ancestor had one or several RPA homologs (Fig. 7), because in several organisms such as *Physcomitrella*, *Chlorella, Acanthamoeba* and *Cryptosporidium* we find both, a core RPA and a homolog that is equally distant from the core RPA as the PHRPA clade. We therefore suspect that the donor of plant Helitron RPA

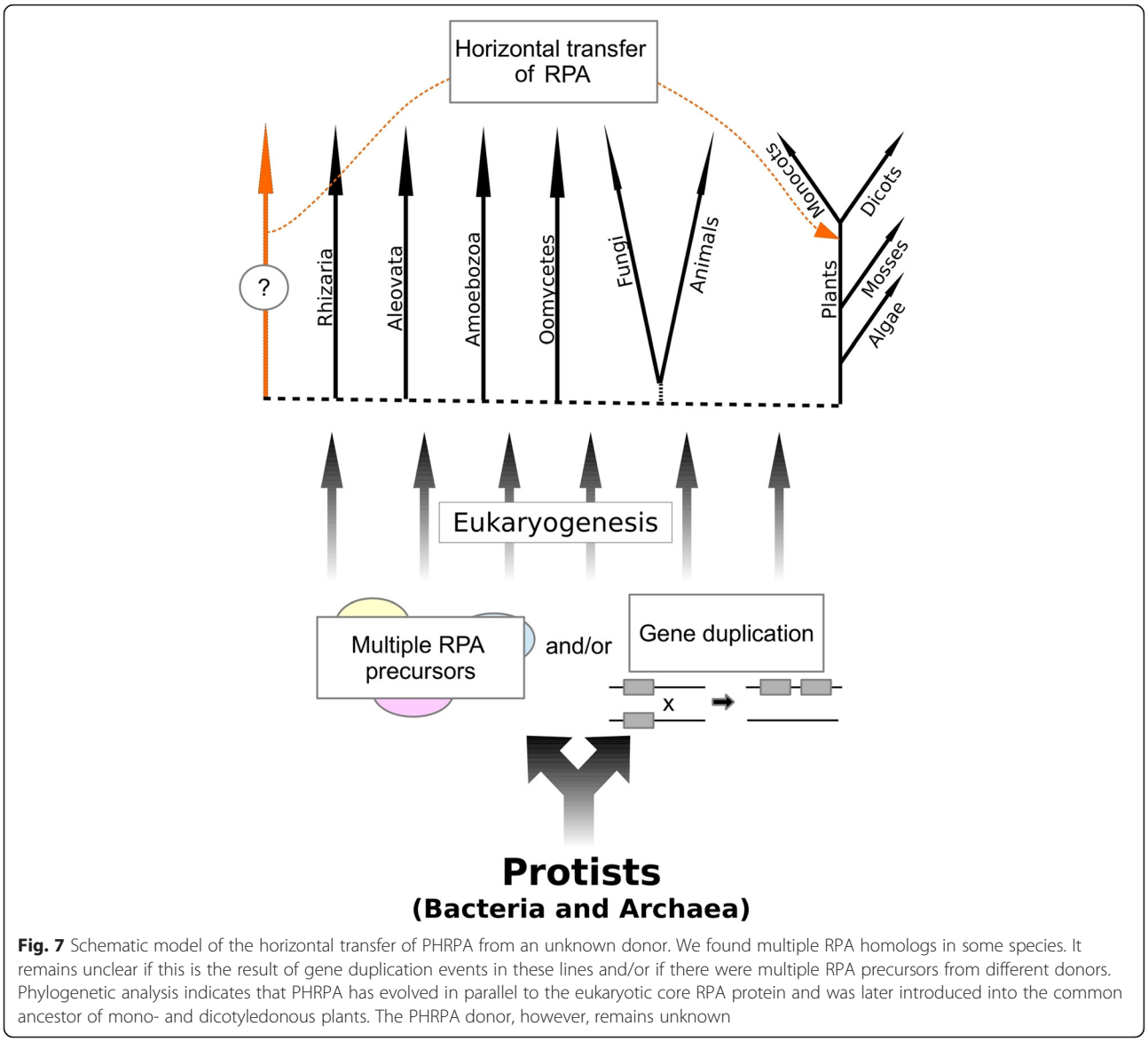


Fig. 7 Schematic model of the horizontal transfer of PHRPA from an unknown donor. We found multiple RPA homologs in some species. It remains unclear if this is the result of gene duplication events in these lines and/or if there were multiple RPA precursors from different donors. Phylogenetic analysis indicates that PHRPA has evolved in parallel to the eukaryotic core RPA protein and was later introduced into the common ancestor of mono- and dicotyledonous plants. The PHRPA donor, however, remains unknown

homologs was probably a basal eukaryote similar to those mentioned above.

In Prokaryotes (bacteria and archaea), horizontal gene transfer is common and and it is believed to be a major mechanism for adaptation [39]. It becomes more and more evident that horizontal transfer is also a common process in eukaryotes. For example the extremophilic red alga *Galdieria sulphuraria* exhibits a enormous metabolic flexibility which it acquired by various genes from different bacteria and archaea [40]. Like genes, also TEs (if they are not the vector for gene transfer themselves) can be transfered between hosts. Often this involves intermediate vectors such as blood feeding insects or pathogens carrying bacteria or viruses to their new hosts. For example in 24 species of the insect order Lepidoptera two non-autonomous *Helitrons* were identified which were also found in the genomes of several double-stranded DNA polydnaviruses [41]. In plants, up to two million horizontal TE transfers only of LTR-retrotransposons were suggested by a comparative analysis among flowering plants [42].

However, what makes the case of PHRPA special is that the proposed horizontal transfer resulted in a successful new type of TE whose widespread distribution in monocots and dicots suggests advantages over normal Helitrons lacking this gene. Indeed, Dong et al. [43] described how stepwise acquisition of gene fragments can produce elements of increasing complexity.

Interestingly, our analysis also suggests that RPA homologs in *Drosophila*, called Helentrons, might also have been acquired though horizontal transfer. But the phylogenetic analysis indicates that they are of an even more

distant origin. Furthermore, highly divergent RPA homologs were also found in Helitrons of zebrafish and starlet sea anemone [12]. However, here we were not able to identify any homology to PHRPAs, which is why they were not included in our phylogenetic analysis. Thus, it appears that Helitrons acquired single-strand binding proteins at least three times independently during evolution, suggesting convergent evolution.

A model for the evolution of semi-autonomous and non-autonomous plant Helitrons

Our data suggest that the numerous non-autonomous *Mothra* elements are mobilized by a single mother element. Surprisingly, this putative mother element encodes for PHRPA but not for a RepHel protein. We speculate that the mother element might itself be depending on a related and fully autonomous element. Indeed, we found one candidate Helitron family that shows strong homology with the RPA protein and the termini of the *Mothra* mother element. We referred to that Helitron family as the *Mothra* sister-family.

Based on these observations, we propose a model which introduces the putative mother element as an additional level of "semi-autonomy" (Fig. 8). We assume that the ancient Helitron consisted of a *RepHel* gene and probably the structural features like the 3' hairpin that we find to be common in all Helitrons. According to our model, the PHRPA protein was then introduced in the common ancestor of mono- and dicots via horizontal transfer 145–300 mya [32, 33] where it got acquired by the progenitor of all RPA containing plant Helitrons (discussed above). We propose that at a later point, one Helitron lineage lost its *RepHel* gene, resulting in the putative *Mothra* mother element that only contains the *PHRPA* gene. This semi-autonomous element would still fulfill some functions in the transposition process but would rely

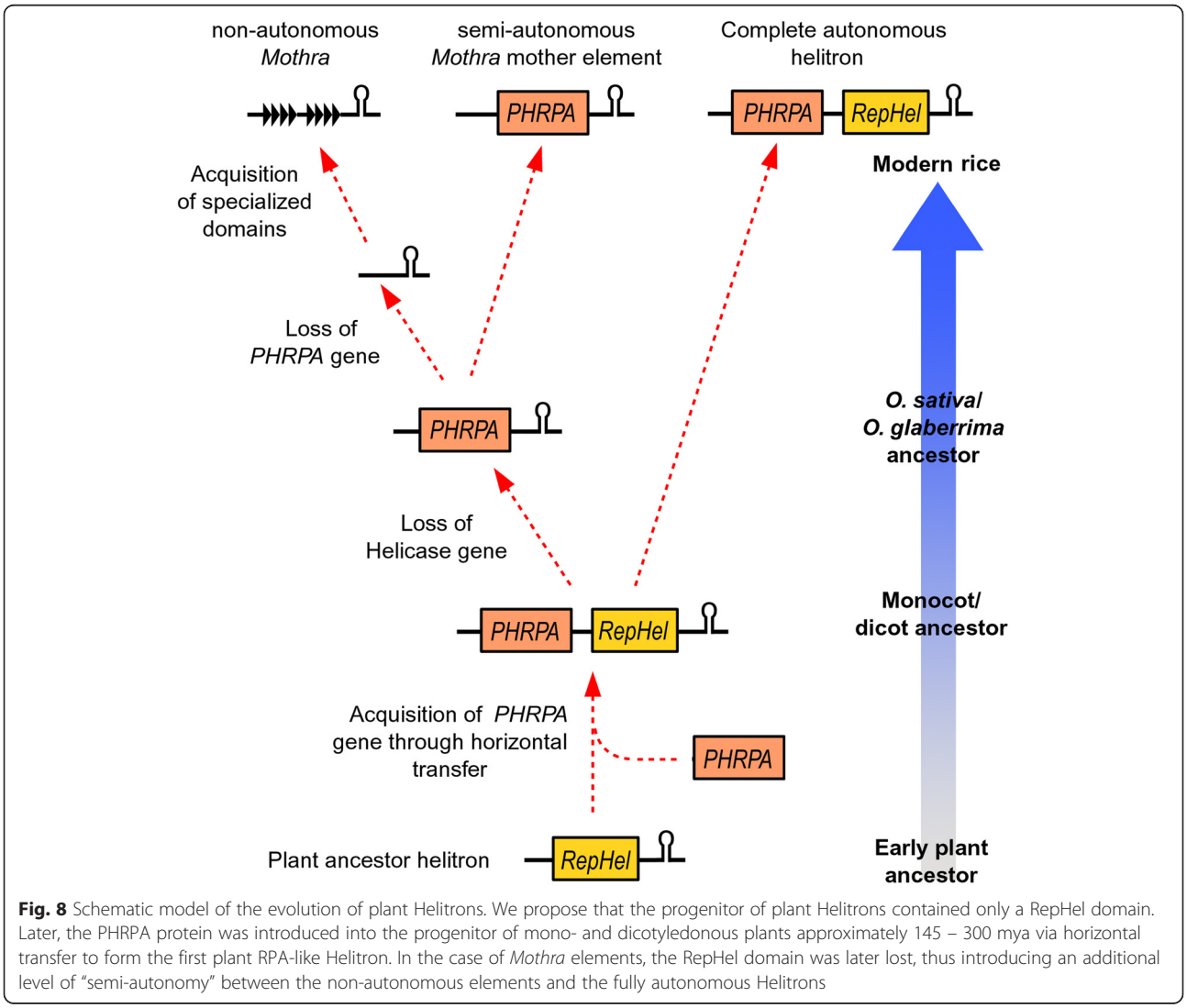


Fig. 8 Schematic model of the evolution of plant Helitrons. We propose that the progenitor of plant Helitrons contained only a RepHel domain. Later, the PHRPA protein was introduced into the progenitor of mono- and dicotyledonous plants approximately 145 – 300 mya via horizontal transfer to form the first plant RPA-like Helitron. In the case of *Mothra* elements, the RepHel domain was later lost, thus introducing an additional level of "semi-autonomy" between the non-autonomous elements and the fully autonomous Helitrons

on the RepHel protein provided by the *Mothra* sister-family. Loss of internal sequences is common during transposition of Helitrons [43]. Furthermore, the evolution of non-autonomous transposable elements has been described in virtually all TE superfamilies [16].

According to our model, the next step in *Mothra* evolution was the loss of the *PHRPA* gene, resulting in a completely non-autonomous element that relies both on the *Mothra* mother element and functional copies of the *Mothra* sister-family (Fig. 8). Finally, the non-autonomous Mothra element acquired the complex tandem repeat blocs which, we propose, improved its transposition efficiency. This proposed stepwise evolution ultimately led to the situation we find in modern rice species where all three types of elements (fully autonomous, semi-autonomous and non-autonomous) exist side-by-side. However, biochemical assays will be needed to confirm the functional relationship between the described elements.

Conclusion

Analysis of the *Mothra* family of Helitrons has provided unexpected insight in to the early evolution of plant Helitrons through the identification of a putative horizontal gene transfer that resulted in a successful sub-group of the Helitron superfamily. Furthermore, the great success of the non-autonomous *Mothra* elements suggests that combinations of different levels of transposition autonomy might be particularly efficient in Helitrons.

Methods

Mothra annotation

To generate the *Mothra* consensus sequence, we extracted and aligned 100 putative copies including 5 kb of flanking sequence which we used to manually determine the boundaries of the element. The identified termini matched the previously described canonical Helitron termini [16]. To deduce the consensus sequences for the sub-types and finally the consensus sequence of the non-autonomous *Mothra* element, we used the multiple alignment software Clustal X [44], the graphical dot-matrix program Dotter from the SeqTools package (https://www.sanger.ac.uk/resources/software/seqtools/) and in-house Perl scripts which are available upon request. To annotate *Mothra* elements we used the *Mothra* consensus sequence in Blastn searches against the *O. sativa Nipponbare* cultivar genome (Version 5) provided by the International Rice Genome Sequencing Project (IRGSP) (plantbiology.msu.edu/pub/data/) [45]. We included hits with a minimum length of 80 basepairs and at least 80 % identity. Because we found many fragments, we merged all hits that were found within 200 bps of flanking sequence to single hits.

To identify the *Mothra* mother element we used Blastn searches of the first 50 and the last 80 bps of the *Mothra* consensus sequence. We considered fragments where we found both ends in the same orientation and that were located within 25 kb from each other. We used the online NCBI platform (http://blast.ncbi.nlm.nih.gov/Blast.cgi) to perform Blastn and Blastx searches against the 323 putative sequences to identify the *RPA* gene. To identify the polymorphisms between *O. sativa* and *O. glaberrima* we used the whole genome alignment produced in a previous study [27].

Phylogenetic tree

The sequences for the phylogenetic tree were retrieved from the NCBI database (http://www.ncbi.nlm.nih.gov/). We used the sequences of the identified Mothra RPA and the core RPA of *O. sativa* as queries and searched each of the main eukaryotic groups, animal, fungi, plants, Alveolata, Amoebae, Rhizaria, Oomycetes and archaea separately. We aligned them using Clustal X [44] with the following parameters for multiple alignments: Gap opening penalty of 10 and Gap extension penalty of 0.1. The phylogenetic tree was generated using MrBayes 3.2.2 [46]. We conducted two runs with 4 chains, each for 4 million generations, sampling every 500 generations. We used all the protein models available in MrBayes and used a reversible jump Monte Carlo Markov Chain (MCMC) [47]. Heterogeneity of substitution rates among different sites was modeled with a gamma distribution. The first quartile of generations was discarded (burn-in) and convergence was evaluated with the average standard deviation of split frequencies (0.002). To illustrate and re-root the tree we used the program Figtree (http://tree.bio.ed.ac.uk/software/figtree/).

Data access

Sequences of *Mothra* elements were deposited in the TREP database (http://www.botinst.uzh.ch/research/genetics/thomasWicker/TREP.html). Sequence alignments that were used for phylogenetic analyses as well as in-house Perl scripts are available upon request.

Additional files

Additional file 1: Figure S1. Distribution of identities between plant helitron RPAs and eukaryotic "core" RPAs. (PDF 274 kb)

Additional file 2: Table S1. Transcripts that were identified, encoding the *Mothra PHRPA* gene and its sister-family *PHRPA* and *RepHel* genes, respectively. (PDF 42 kb)

Abbreviations

TE: Transposable element; TIR: Terminal inverted repeat; aa: Amino acid; RPA: Replication protein A; ORF: Open reading frame; PHRPA: Plant helitron replication protein A; mya: Million years ago.

Competing interests
The authors declare that they have no competing interests.

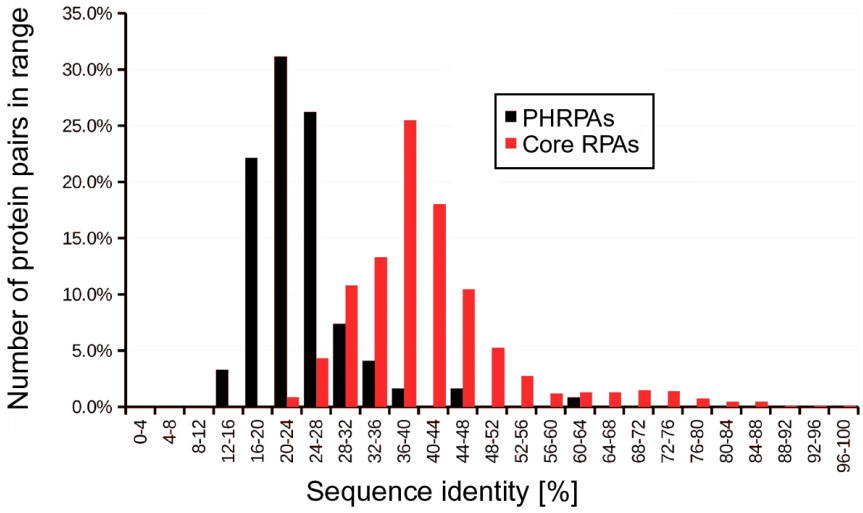
Authors' contributions
SR performed the analysis and wrote the paper. TW designed the study and wrote the paper. FM helped with the phylogenetic analysis. All authors have read and approved the final version of the paper.

Acknowledgements
This study was supported by the Swiss National Foundation grant # 31003A_138505/1.

Received: 24 September 2015 Accepted: 4 December 2015
Published online: 17 December 2015

References

1. Kapitonov VV, Jurka J. Rolling-circle transposons in eukaryotes. PNAS. 2001;98(15):8714–9.
2. Yang L, Bennetzen JL. Structure-based discovery and description of plant and animal Helitrons. PNAS. 2009;106(31):12832–7.
3. Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, et al. The *Physcomitrella* Genome Reveals Evolutionary Insights into the Conquest of Land by Plants. Science. 2008;319:64–9.
4. Hood ME. Repetitive DNA in the automictic fungus *Microbotryum violaceum*. Genetica. 2005;124(1):1–10.
5. Poulter RTM, Goodwin TJD, Butler MI. Vertebrate helentrons and other novel Helitrons. Gene. 2003;313:201–12.
6. Nierman WC, Pain A, Anderson MJ, Wortman JR, Kim HS, Arroyo J, et al. Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. Nature. 2005;438:1151–6.
7. Kapitonov VV, Jurka J. Helitron-1_SP, a family of autonomous Helitrons in the sea urchin genome. Repbase Rep. 2005;5:393.
8. Zhou Q , Froschauer A, Schultheis C, Schmidt C, Bienert GP, Wenning M, etal. Helitron transposons on the sex chromosomes of the platyfish *Xiphophorus maculatus* and their evolution in animal *genomes*. Zebrafish. 2006;3:39–52.
9. Ennio C, De Iorio S, Capriglione T. Identification of a novel helitron transposon in the genome of Antarctic fish. Mol Phylogenet Evol. 2011;58(3):439–46.
10. Pritham EJ, Feschotte C. Massive amplification of rolling-circle transposons in the lineage of the bat *Myotis lucifugus*. PNAS. 2007;104(6):1895–900.
11. Yang L, Bennetzten JL. Distribution, diversity, evolution, and survival of *Helitrons* in the maize genome. PNAS. 2009;106(47):19922–7.
12. Kapitonov VV, Jurka J. *Helitrons* on a roll: eukaryotic rolling-circle transposons. Science. 2007;23(10):521–9.
13. Chandler M, de la Cruz F, Dyda F, Hickman AB, Moncalian G, Ton-Hoang B. Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. Nat Rev Micro. 2013;11(8):525–38.
14. Li Y, Dooner HK. Excision of Helitron transposons in maize. Genetics. 2009;182(1):399–402.
15. Thomas J, Pritham EJ. Helitrons, the eukaryotic rolling-circle transposable elements. Microbiol Spectrum. 2015;3(4):MDNA3-0049-2014.
16. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2003;8:973–82.
17. Stenger DC, Revington GN, Stevenson MC, Bisaro DM. Replicational release of geminivirus genomes from tandemly repeated copies: evidence for rolling-circle replication of a plant viral DNA. PNAS. 1991;88(18):8029–33.
18. Mendiola MV, Bernales I, De La Cruz F. Differential roles of the transposon termini in IS91 transposition. PNAS. 1994;91(5):1922–6.
19. Oakley GG, Patrick SM. Replication protein A: directing traffic at the intersection of replication and repair. Front Biosci. 2010;15:883.
20. Wold MS. Replication protein A: a heterotrimeric, single-stranded DNA-binding protein required for eukaryotic DNA metabolism. Annu Rev Biochem. 1997;66(1):61–92.
21. Thomas J, Vadnagara K, Pritham EJ. DINE-1, the highest copy number repeats in Drosophila melanogaster are non-autonomous endo-nuclease-encoding rolling-circle transposable elements (Helentrons). Mob DNA. 2014;5:18.
22. Lai J, Li Y, Messing J, Dooner HK. Gene movement by Helitron transposons contributes to the haplotype variability of maize. PNAS. 2005;102(25):9068–73.
23. Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet. 2005;37(9):997–1002.
24. Barbaglia AM, Klusman KM, Higgins J, Shaw JR, Hannah LC, Lal SK. Gene capture by Helitron transposons reshuffles the transcriptome of maize. Genetics. 2012;190(3):965–75.
25. Han MJ, Shen YH, Xu MS, Liang HY, Zhang HH, Zhang Z. Identification and Evolution of the Silkworm Helitrons and their Contribution to Transcripts. DNA Res. 2013;20:471–84.
26. Thomas J, Phillips CD, Baker RJ, Pritham EJ. Rolling-Circle Transposons Catalyze Genomic Innovation in a Mammalian Lineage. Genome Biol Evol. 2014;6(10):2595–610.
27. Roffler S, Wicker T. Genome-wide comparison of Asian and African rice reveals high recent activity of DNA transposons. Mob DNA. 2015;6(1):8.
28. Kemetaka S, Okano T, Ohsumi M, Ohsumi Y. Apg14p and Apg6/Vps30p Form a Protein Complex Essential for Autophagy in the Yeast, *Saccheromyces cerevisiae*. J Biol Chem. 1998;273(35):22284–91.
29. Wang M, Yu Y, Haberer G, Marri PR, Fan C, Goicoechea JL, et al. The genome sequence of African rice (Oryza glaberrima) and evidence for independent domestication. Nat Genet. 2014;46:982–8.
30. Kikuchi K, Terauchi K, Wada M, Hirano HY. The plant MITE mPing is mobilized in anther culture. Nature. 2003;421:167–70.
31. Koonin EV. The origin and early evolution of eukaryotes in the light of phylogenomics. Genome Biol. 2010;11(5):209.
32. Kawai Y, Otsuka J. The deep phylogeny of land plants inferred from a full analysis of nucleotide base changes in terms of mutation and selection. J Mol Evol. 2004;58:479–89.
33. Zimmer A, Lang D, Richardt S, Frank W, Reski R, Rensing SA. Dating the early evolution of plants: detection and molecular clock analyses of orthologs. Mol Genet Genomics. 2007;278:393–402.
34. Buchmann JP, Matsumoto T, Stein N, Keller B, Wicker T. Interspecies sequence comparison of Brachypodium reveals how transposon activity corrodes genome colinearity. Plant J. 2012;488:213–7.
35. Yang G, Weil CF, Wessler SR. A rice Tc1/mariner-like element transposes in yeast. Plant Cell. 2006;18:2469–78.
36. Yang G, Nagel DH, Feschotte C, Hancock CN, Wessler SR. Tuned for transposition: Molecular determinants underlying the hyperactivity of a *Stowaway* MITE. Science. 2009;325(5946):1391–4.
37. Lisch D. Epigenetic regulation of transposable elements in plants. Plant Biol. 2009;60:43–66.
38. Feschotte C, Wessler SR. Treasures in the attic: rolling circle transposition discovered in eukaryotic genomes. PNAS. 2001;98(16):8923–4.
39. Rocha EPC. With a little help from prokaryotes. Science. 2013;339(6124):1154–5.
40. Schönknecht G, Chen WH, Ternes CM, Barbier GG, Shrestha RP, Stanke M, et al. Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. Science. 2013;339(6124):1207–10.
41. Coates BS. Horizontal transfer of a non-autonomous Helitron among insect and viral genomes. BMC Genomics. 2015;16(1):137.
42. El Baidouri M, Carpentier MC, Cooke R, Gao D, Lasserre E, Llauro C, et al. Widespread and frequent horizontal transfers of transposable elements in plants. Genome Res. 2014;24(5):831–8.
43. Dong Y, Lu X, Song W, Shi L, Zhang M, Zhao H, et al. Structural characterization of helitrons and their stepwise capturing of gene fragments in the maize genome. BMC Genomics. 2011;12:609.
44. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23:2947–8.
45. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. Nature. 2005;436:793–800.
46. Ronquist F, Teslenko M, van der Mark P, Ayers DL, Darling A, Höhna S, et al. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Syst Biol. 2012;61(3):539–42.
47. Huelsenbeck JP, Larget B, Alfaro ME. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. Mol Biol Evol. 2004;21(6):1123–33.



Additional Figure 1. Levels of sequence identity of RPA core and PHRPA proteins. For this analysis, all proteins within the two groups were compared pairwise. The x-axis shows the degree of sequence identity at the protein level while the y-axis shows the percentage of protein pairs in each class.

Additional Table 1. Transcripts of genes from the Mothra mother element and its sister-family.

Gene	GenBank accession
Mothra PHRPA	HS381803
Sister-family *PHRPA*	NM_001187565
Sister-family *RepHel*	CF306916
Sister-family *RepHel*	CF305553

Chapter 5:

DNA transposons specifically accelerate evolution of genes in rice and other grasses

This chapter describes how DNA transposon activity influences regulatory regions but also coding sequences of genes in grasses indirectly through DSB-repair. This suggests DNA TEs as a major driver of grass evolution. This is currently in submission (Wicker *et al. 2016*).

DNA transposons specifically accelerate evolution of genes in rice and other grasses

Thomas Wicker[1,11], Yeisoo Yu[2,10], Georg Haberer[3], Klaus F X Mayer[3], Pradeep Reddy Marri[4], Steve Rounsley[4], Mingsheng Chen[5], Andrea Zuccolo[6], Olivier Panaud[7], Rod A Wing[2,8,9], & Stefan Roffler[1]

[1]Department of Plant and Microbial Biology, University of Zurich, Switzerland.
[2]Arizona Genomics Institute, School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA
[3]Plant Genome and Systems Biology, Helmholtz Center Munich, Neuherberg, Germany.
[4]Dow AgroSciences, Indianapolis, Indiana, USA.
[5]State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China.
[6]Institute of Life Sciences, Scuola Superiore Sant'Anna, Pisa, Italy.
[7]Laboratoire Génome et Développement des Plantes, UMR5096 UPVD/CNRS. Université de Perpignan Via Domitia. 66860 Perpignan France.
[8]International Rice Research Institute, Los Baños, Philippines,
[9]Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, USA.
[10]Present addresses - Phyzen Genomics Institute, Phyzen Inc., Seoul, South Korea, 151-836.
[11]Corresponding author.

Corresponding author's contact information:

Thomas Wicker

Institute of Plant Biology, University of Zurich, Zollikerstrasse 107, CH-8008 Zurich, Switzerland

Email: wicker@botinst.uzh.ch

Keywords: double-strand break repair, DNA transposon, excision, error-prone DNA repair

Abstract

DNA (Class 2) transposons are mobile genetic elements which move within their "host" genome through excising and re-inserting elsewhere. Although the rice genome contains tens of thousands of such elements, their actual role in evolution is still unclear. Here, we analyzed over 650 transposon polymorphisms in the rice species *Oriza sativa* and *O. glaberrima.* Interestingly, we found that transposon excisions usually go along with the introduction of numerous mutations in the sequences neighboring the transposon. We found that the 3,000 bp flanking the excised transposons can contain over 10 times more mutations than the genome-wide average. The observed mutation patterns are consistent with products of double-strand break induced DNA replication and DNA translesion synthesis which are highly error-prone (1-4). Because DNA transposons preferably insert near genes (5-7), their excisions specifically increase substitution rates in coding sequences and regulatory regions. Most importantly, we found this phenomenon also in maize, wheat and barely, indicating that DNA transposons accelerate gene evolution in the entire grass (Poaceae) family. Thus, these findings identify DNA transposons as a major evolutionary force in gene evolution in the grass family which contains over 10,000 species and includes the most important agricultural crops.

Main text

Transposons excisions leave double-strand breaks (DSBs) that have to be repaired by the cell. Depending on the repair pathway, this can lead to deletions and/or insertions of "filler" sequences at the site of the DSB (8-10). Sometimes, re-arrangements at the excision site can be so extensive that excisions are difficult to identify (9,10, Supplementary note A). Considering the complex DSB repair processes, we wanted to study DNA repair patterns at excision sites at the genome-wide level using the closely related rice species *O. sativa* and *O. glaberrima* which diverged approximately 600,000 years ago (11).

For our analysis, we annotated 27,641 DNA transposons in the *O. sativa* genome. They show a strong preference to insert close to transcription start and end points of genes (Supplementary Fig. 1), which is in agreement with previous findings (5-7, Supplementary note B). To identify DNA transposon polymorphisms, we compared the annotated transposons loci with their orthologs in *O. glaberrima.* We manually screened over 2,000 potential polymorphisms and classified 482 as insertions and 158 as excisions (Methods, Supplementary Tables 1 and 2, Supplementary note C). Here, we made particular efforts to ensure that indeed orthologous loci were compared (methods, Supplementary Fig. 2, Supplementary note D). The polymorphic transposons belong to five different superfamilies of which *DTT_Mariner* and *DTH_Harbinger* elements comprise the majority (Supplementary Table 3).

Interestingly, we found that excisions often go along with the introduction of numerous nucleotide substitutions and small insertions and deletions (InDels) in sequences flanking the transposons, with some flanking region containing over 10 times more mutations than the genome on average (example in Fig. 1). To quantify this effect, we analyzed the 12 kb flanking each polymorphic transposon and added up all nucleotide substitutions and (InDels) relative to the transposon insertion/excision site. The resulting plot shows that the overall frequency of nucleotide substitutions and InDels increases in an exponential manner towards the TE excision site to at least four-fold on average, compared to randomly picked genomic sequences (Fig. 2). Numbers of nucleotide substitutions and InDels are on increased up to a distance of 3 kb from the excision point (Fig. 2). In contrast, transposon insertion sites have many fewer mutations in flanking regions, showing only a small increase in nucleotide substitution frequency in their neighborhood (Fig. 1, Supplementary Fig. 3).

Considering findings on DSB repair from yeast (1-4, 12), we propose a molecular mechanism that explains the high numbers of mutations flanking transposon excisions in rice (Fig. 2C): In the first step, the transposons excises from the genome, leaving a DSB for the cell to repair. After transposon excision, 3' overhangs are produced by exonucleases (Fig. 4, step 1). The 3' overhangs then anneal using micro-homologies of a few bp (Fig. 4, step 2), or through an intermediate generated by invasion of a foreign

strand (Supplementary note A). Subsequently, the single-stranded DNA segments are used as templates for the synthesis of a new second strand, which is the step that introduces numerous mutations (Fig. 2C, step 3). We propose that DNA replication is analogous to that described for DSB-induced replication in yeast. Here, mutations are introduced by translesion synthesis involving DNA polymerase zeta (2) and by a DSB-induced replication complex that has deficiencies in DNA polymerase delta fidelity and mismatch repair (3). Possibly, Rev1 polymerase also contributes to erroneous DNA repair (4). The end product of the repair process are sequence segments flanking the transposon excision which are riddled with nucleotide substitutions and small InDels (Fig. 2C, step 4 and Fig. 1). The length of the segment containing the mutations depends on the size of the 3' overhang produced in the initial repair step. In yeast, these overhangs can be several kb in size (1) and our data indicate this to be similar in rice, since the average nucleotide substitution frequency levels off approximately 3 kb away from the excision site (see Fig. 2A and 2B).

Because DNA transposons preferably reside in gene promoters, we expected that these regions should evolve at a particularly high rate. Indeed, we found that the 2000 bp upstream of genes consistently contain 20-29% more nucleotide substitutions than intergenic sequences from the same chromosomal region (Fig. 3, Supplementary Table 4). Because the genomes of the closely related *O. sativa* and *O. glaberrima* are ~99.5% identical on average, the differences in sequence conservation between promoters and intergenic sequences are small, but the large sample size assures that they are highly significant (P-value <2.2E-16). Intergenic regions in rice are mostly comprised of Class 1 retrotransposons which are believed to be largely free from selection pressure. It is therefore intriguing that transposon activity apparently increases the mutation rate of promoters to a degree that they evolve more rapidly than selectively neutral sequences. Interestingly, sequence conservation is generally lower in the centromeric and pericentromeric regions of chromosomes than in distal regions (Fig. 3), for which we have no explanation at this point.

The preference of DNA transposons to reside in up- and downstream regions of genes also implied that

the 5' and 3' ends of coding sequences (CDS) should show an overall higher substitution rate than their central parts. Thus, we aligned CDS of closest homologs from *O. sativa* and *O. glaberrima* and studied overall sequence conservation as well as distributions of nucleotide substitutions along the aligned CDS. Overall, most CDS from *O. sativa* and *O. glaberrima* CDS are over 99.5% identical. However, the distribution of sequence identities trails off with some CDS being less than 97% identical (Supplementary Fig. 4). We expected that CDS which are >99.5% identical have not experienced transposon excisions in their vicinity, while genes with lower sequence identity could be those that have accumulated mutations due to a nearby transposon excisions. Indeed, we found that genes with lower than median sequence identity ranging from 98% to 99.4% show a >27% higher number of substitutions in their 5' and 3' regions than in the central part of the CDS (Fig. 4A, Supplementary Table 5), while genes with higher levels of sequence conservation do not show this pattern (Supplementary Fig. 5). Here, we only considered nucleotide substitutions in synonymous sites to avoid effects of differing selection pressures in different parts of the genes.

Because all grass genomes sequenced so far are rich in DNA transposons, we predicted that we would find this phenomenon also in other grasses. We therefore compared closest gene homologs from wheat and barely, two species which diverged approximately 8 Myr ago (13). Indeed, the 5' and 3' regions of the genes show a more than 20% higher number of substitutions than the central part of the genes (Fig. 4B). We also analyzed maize where many genes are present in duplicates because maize is a relatively young polyploid that underwent a whole-genome duplication 5-10 Myr ago (14). Thus, a comparison of such intra-genomic closest homologs is analogous to a comparison of genes between two species. Here we found an even stronger effect, with 5' and 3' regions showing almost 30% more substitutions than the central part of genes (Fig. 4C). For both, the wheat/barely and the maize intra-genomic CDS comparisons, the effects are statistically highly significantly (Supplementary Table 5). Considering that rice, maize, wheat and barley represent three different major clades of the grasses, our data strongly indicate that the described higher mutation rates in genes and regulatory sequences is common to all

grasses. Interestingly, we did not find elevated mutation rates in genes in representatives of dicotyledonous plants ("dicots") such as *Arabidopsis*, *Brassica*, poplar and soybean (example in Fig 4D, Supplementary Fig. 6, Supplementary note E). A *de novo* search for class 2 elements in these dicot genomes revealed that they contain at least 100 times fewer DNA transposons than grasses (Supplementary Fig. 7, Supplementary note E). Thus, this strengthens the correlation even more between the presence of DNA transposons and increased mutation rates of genes.

Data on how TEs contribute to gene evolution has been somewhat anecdotal (examples in 15-17). So far, most widely accepted is their role in altering gene expression. For example, an TE-mediated increase in expression level of the *tb1* gene in maize resulted in plants with fewer branches, a fundamental step in maize domestication (18,19). We did indeed find that the presence of transposons is associated with higher levels of DNA methylation sites, suggesting an effect on transcription (Supplementary Fig. 8, Supplementary note G). However, they main contrast to previous studies is that our data show that transposon activity can directly change coding sequences and regulatory regions by introducing nucleotide substitutions and InDels. Furthermore, previous studies showed that transposon excisions can cause deletions and insertions of filler sequence at the immediate site of the excisions (8-10). We now found that, in addition to these local rearrangements at the very site of the excision, the repair process can introduce many mutations hundreds or even thousands of bp away from the excision site. This has the profound result that, even if the excision changes only a few bp at the actual transposon site, the entire genomic region accumulates mutations as a result of error-prone strand synthesis (see Fig. 4, Supplementary note F). Most importantly, we could show that this affects thousands of genes in the species studied, and we provide evidence that this phenomenon is common to the vast family of the grasses with its over 10,000 species. Our data thereby also indicate that the highly successful types of non-autonomous DNA transposons elements that drive the accelerated evolution of genes only evolved after the separation of monocotyledons and dicotyledons approximately 145-300 Myr ago (20,21). We previously showed that about 3% of the DNA transposons in rice have moved within the past 600,000

years, indicating that these elements are highly active (10). Since DNA transposons are present in tens of thousands of copies in grasses (5,14), most genes will experience transposon excisions in their proximity at some point and therefore accumulate particularly high numbers of mutations over time. Consequently, we found a stronger mutation rate gradient in more distantly related grasses such as wheat and barley (see Fig. 4). We conclude that the activity of DNA transposons is a major driving force in the evolution of grasses, because they specifically accelerate evolution of genes. Our findings may, in part, explain the phenomenal evolutionary success of the grasses, a very large group of plants that contains the most important crops such as rice, maize, wheat, sorghum and barley which are the basis of most food consumed by mankind.

Acknowledgements

We thank Detlef Weigel and Claude Becker for providing methylome data for *O. sativa* and *O. glaberrima*. This material is based on work supported by the Swiss National Foundation grant #31003A_138505/1 to T.W., by the US National Science Foundation under grants #0321678, #0638541, #0822284 and #1026200 to Y.Y., and R.A.W. and the Bud Antle Endowed Chair of Excellence in Agriculture and Life Sciences and the AXA Endowed Chair of Genome Biology and Evolutionary Genomics to R.A.W. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the US National Science Foundation.

Competing interests

The authors declare no conflicts of interest.

Author contributions

TW designed the study, analyzed the data and wrote the paper. SR helped design the study, created

software, analyzed the data and provided critical input in writing the manuscript. RAW and coworkers produced the genome sequence.

References

1. Storici F, Snipe JR, Chan GK., Gordenin, D.A. Resnick, M.A. Conservative repair of a chromosomal double-strand break by single-strand DNA through two steps of annealing. J Cell Biol. 2006; **26:** 7645-7657.

2. Yang, Y, Sterling, J, Storici, F, Resnick, MA and Gordenin, DA. Hypermutability of damaged single-strand DNA formed at double-strand breaks and uncapped telomeres in yeast Saccharomyces cerevisiae. PLoS Genetics 2008; **4:** e1000264.

3. Deem A, Keszthelyi A, Blackgrove T, Vayl A, Coffey B, Mathur R, Chabes A, Malkova A. Break-induced replication is highly inaccurate. PLoS Biol. 2001; **9:** e1000594.

4. Nakagawa M, Takahashi S, Narumi I, Sakamoto AN. Role of AtPolζ, AtRev1 and AtPolη in γ ray-induced mutagenesis. Plant Sigal Behav. 2011; **26:** 728-731.

5. International *Brachypodium* Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature. 2010; **463:** 763–768.

6. Bureau T, Wessler SR. Mobile inverted-repeat elements of the Tourist family are associated with the genes of many cereal grasses. Proc Natl Acad Sci USA 1994; **9:** 907-916.

7. Bureau T, Wessler SR. Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Proc Natl Acad Sci USA 1994; **9:** 1411-1115.

8. Yang G., Weil CF. and Wessler SR. A rice Tc1/mariner-like element transposes in yeast. Plant Cell. 2006; **18:** 2469–2478.

9. Buchmann JP, Matsumoto T, Stein N, Keller B, Wicker T. Interspecies sequence comparison of Brachypodium reveals how transposon activity corrodes genome colinearity. Plant J. 2012; **488:** 213-217.

10. Roffler S, Wicker T. 2015. Genome-wide comparison of Asian and African rice reveals high recent activity of DNA transposons. Mob DNA. 2015; **6:**8.

11. Wang, M., Yu, Y., Haberer, G., Marri, P.R., Fan, C, al. The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. Nat Genet. 2014; **46:** 982-988.

12. Lydeard JR, Jain S, Yamaguchi M, Haber JE. Break-induced replication and telomerase-independent telomere maintenance require Pol32. Nature. 2007; **448:** 820-823.

13. Middleton CP, Senerchia N, Stein N, Akhunov ED, Keller B, Wicker T, Kilian B. Sequencing of chloroplast genomes from wheat, barley, rye and their relatives provides a detailed insight into the evolution of the triticeae tribe. PLoS One. 2014; **9:** e85761.

14. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F. et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 2009; **326:** 1112-1115.

15. Fugmann SD, Lee AI, Shockett PE, Villey IJ, Schatz DG. The RAG proteins and V(D)J recombination: complexes, ends, and transposition. Ann Rev Immunol. 2000; **18:** 495-527.

16. Tsiantis, M.A. A transposon in tb1 drove maize domestication. Nat Genet. 2011; **43:** 1048-1050.

17. Naito K, Monden Y, Yasuda K, Saito H, Okumoto Y. mPing: The bursting transposon. *Breeding* Science. 2014; **64:** 109-1014.

18. Tsiantis, M.A. A transposon in tb1 drove maize domestication. Nat Genet. 2011; **43:** 1048-1050.

19. Clark RM, Wagler TN, Quijada P, Doebley JA. A distant upstream enhancer at the maize domestication gene tb1 has pleiotropic effects on plant and inflorescent architecture. Nat Genet. 2006; **38:** 594–597.

20 Kawai Y, Otsuka J. The deep phylogeny of land plants inferred from a full analysis of nucleotide base changes in terms of mutation and selection. J Mol Evol. 2004; **58:** 479–489.

21. Zimmer, A, Lang, D, Richardt, S, Frank, W, Reski, R and Rensing, SA. Dating the early evolution of plants: detection and molecular clock analyses of orthologs. Mol Gen Genomics. 2007; **278:** 393–402.

22. Puchta, H. The repair of double-strand breaks in plants: Mechanisms and consequences for genome evolution. J Exp Botany. 2005; **56:** 1–14.

23. Hartlerode AJ, Scully R. Mechanisms of double-strand break repair in somatic mammalian cells. Biochem J. 2009; **423:** 157–168.

24. Nassif N, Penney J, Pal S, Engels WR, Gloor GB. Efficient copying of nonhomologous sequences from ectopic sites via P-element-induced gap repair. Mol Cell Biol 1994; **14:** 1613–1625.

Figures

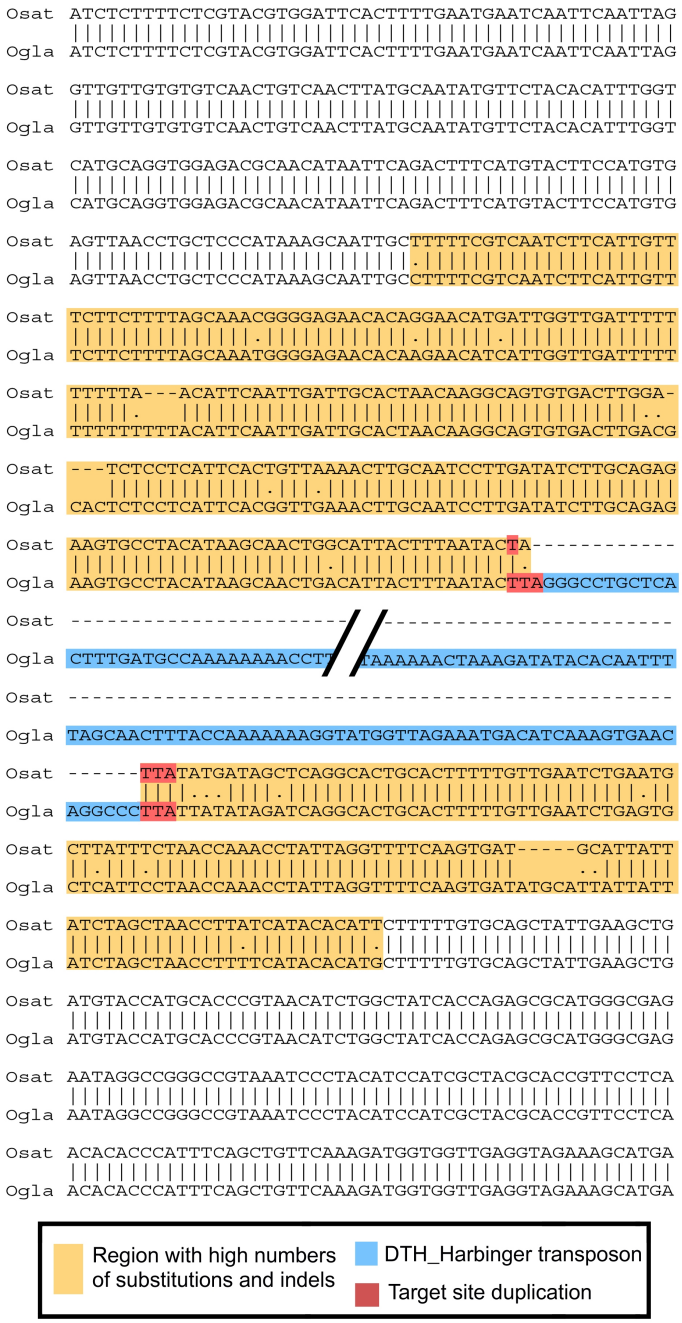


Figure 1. Example for a DNA transposon excision with numerous nucleotide substitutions in its flanking region. A *DTH_Harbinger* transposon excised from the genome of *O. sativa* (Osat) while it remained present in *O. glaberrima* (Ogla). In this particular event, the transposon excised almost perfectly, only losing 2 bp of the target site duplication and replacing one of them with a mis-matching base. The 211 bp upstream and 120 bp downstream of the excision contain 25 mutations (InDels >1bp are counted as one mutation), resulting in less than 93% sequence identity and thus making the mutation rate over 15 times higher than for the genome overall. Outside of the region with the mutations, *O. sativa* and *O. glaberrima* sequences are identical, reflecting the overall genome-wide sequence conservation of ~99.5%. The segments shown correspond to *O. sativa* chromosome 1 position 23,814,561-23,815,081 and *O. glaberrima* chromosome 1 position 16,579,166-16,580,116.

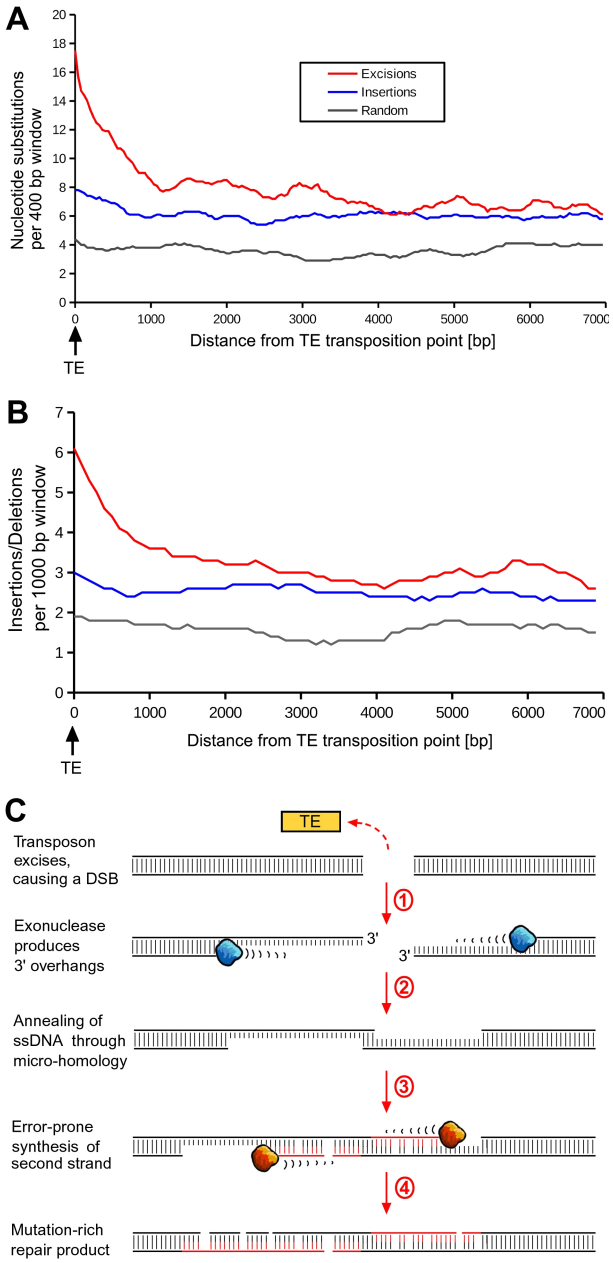


Figure 2. Transposon induced mutations in sequences flanking insertions and excisions. (**A**) Frequencies of nucleotide substitutions and insertions/Deletions (InDels) relative to transposon insertion/excision sites in rice. For the plot, 438 sequence alignments carrying transposon insertions (blue line) and 206 alignments carrying excisions (red line) were compiled. As control, 340 alignments of randomly picked orthologous sequences from *O. sativa* and *O. glaberrima* were used (gray line, see methods). Nucleotide substitution frequency were calculated in a 400 bp sliding window with a 40 bp sliding step. (**B**) InDel frequency calculated in a 1000 bp sliding window with a 100 bp sliding step. (**C**) Proposed mechanism for error-prone DNA repair following the excision of DNA transposons. Step 1: After transposable element (TE) excision, 3' overhangs are generated by exonuclease (blue). Step 2: The 3' overhangs anneal using micro-homologies. To keep it simple, we only represent single-strand annealing (SSA, 22, 23) here. Alternatively, the strands could also be connected via synthesis-dependent strand annealing (SDSA, 22-24), where the two strands are connected by "filler" sequences (which were found in some cases, not shown). Step 3: New strands are synthesized by a replication complex that has deficiencies in DNA polymerases fidelity and mismatch repair. Step 4: The final repair product is rich in nucleotide substitutions and small insertions and deletions.

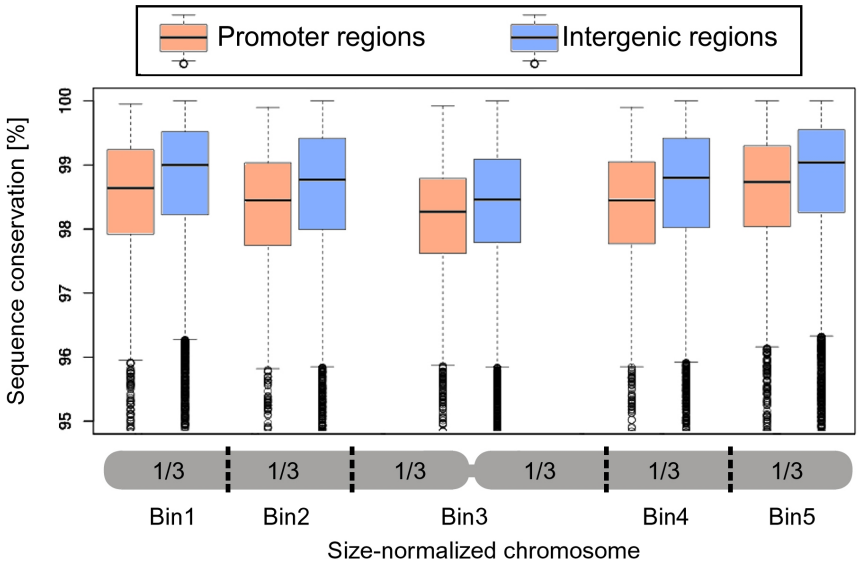


Figure 3. Sequence conservation along *O. sativa* and *O. glaberrima* chromosomes. Data from all 12 chromosomes were compiled and chromosome sizes were normalized by dividing chromosome arms into 3 equally sized bins (x-axis). The y-axis depicts sequence identity of orthologous sequences. For each chromosome bin, promoter regions (the 2,000 bp upstream of the transcription start point, red box plots) are compared with intergenic sequences from the same bin (blue box plots). Promoters are on average 20-29% less conserved than intergenic sequences from the same chromosome bin. To calculate sequence conservation in intergenic regions, we isolated segments that are located in the middle of intergenic sequences which are at least 10 kb in size (i.e. the distance between the end of one gene and the start of the next one is over 10 kb).

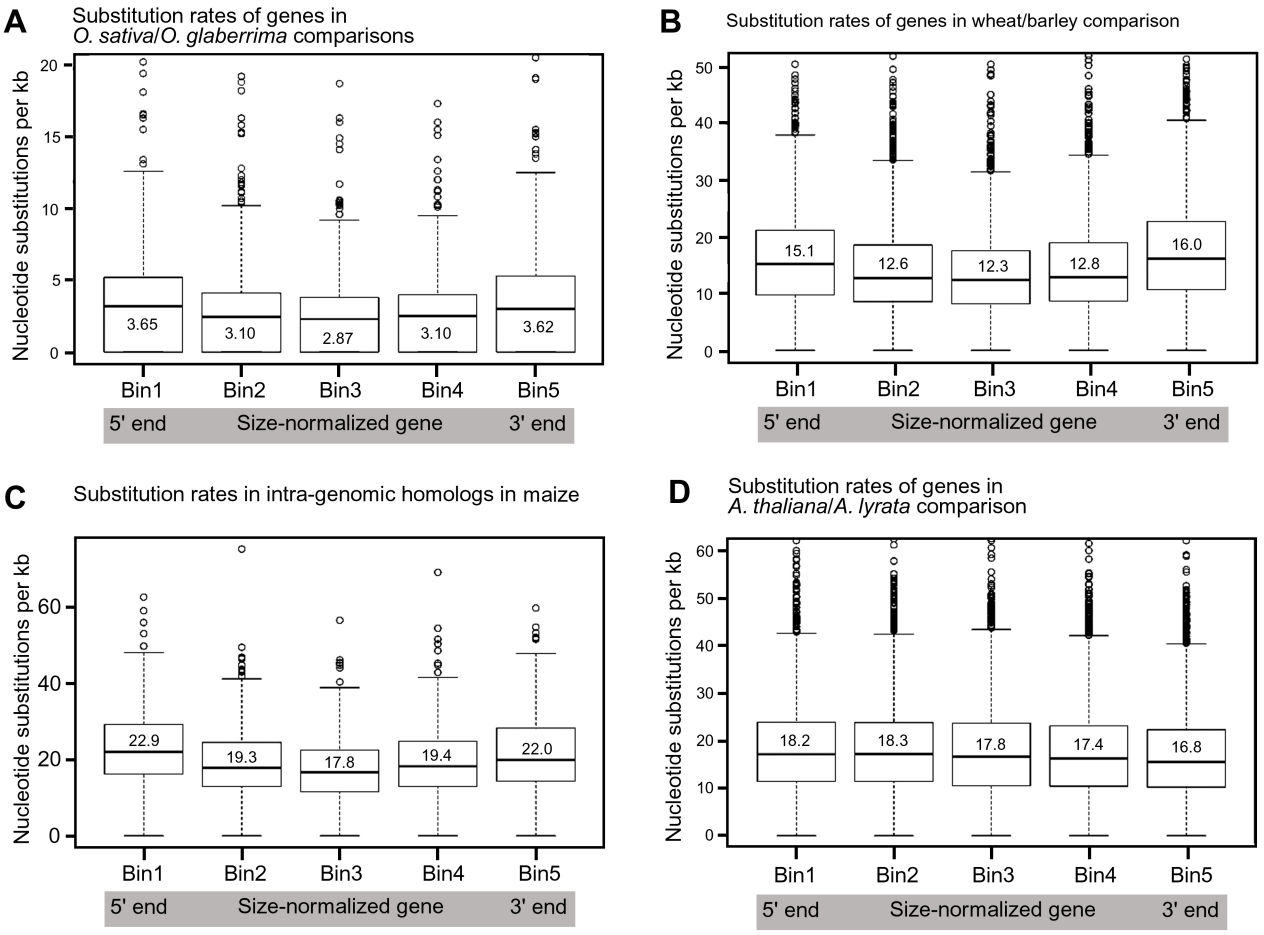


Figure 4. Nucleotide substitution frequencies in synonymous sites of genes showing that genes from grasses have higher mutation rates in their 5' and 3' ends than in the central parts. To normalize the different CDS sizes, genes were divided into 5 equally sized bins and frequencies were normalized to nucleotide substitutions per kb for each bin. The bold line inside the box is the median value, while mean values are indicated with numbers. (**A**) Comparison of 442 closest homologs from *O. sativa* and *O. glaberrima.* (**B**) Comparison of 2,314 pairs of closest homologs from wheat and barley. (**C**) Comparison of 428 pairs of intra-genomic closest homologs in maize that originate from a whole genome duplication. (**D**) Comparison of 4,133 pairs of closest homologs from *A. thaliana* and *A. lyrata*.

Methods

Analysis of the distribution of DNA transposons relative to genes in rice

A total of 101 sequences of *DTT_Mariner* and *DTH_Harbinger* transposons from rice were obtained from the the TREP database (wheat.pw.usda.gov/ggpages/Repeats/). They represent 19 *DTT_Mariner* and 25 *DTH_Harbinger* families. The 101 sequences were mapped with blastn to the *O. sativa* genome (version 6) using an in-house Perl script. The cutoff for blast hits was 50 bp and 80% sequence identity. If multiple transposable element (TE) families mapped to the same location, the one with the strongest blastn hit was chosen. To analyze their position relative to genes, the TE annotation was then cross-matched with the gff format gene annotation of the rice genome. We used the annotated transcription start and end points as anchor points and generated a dataset of the positions of all annotated TEs within 5kb upstream of the transcription start point and 5 kb downstream of the transcription end point for each gene. Furthermore, positions of TEs inside the gene were recorded. We selected genes larger than 4 kb and recorded TE positions within 2 kb from each end of a gene. For simplicity, only genes in forward orientation were used. The final dataset included data for 4,994 genes. Sequences covered by TEs were added up for all genes, resulting in a final coverage plot that reflects the overall distribution of TEs relative to genes (Supplementary Figure 1).

Identification of transposon polymorphisms

We used an alignment of approximately 60% of the genomes of *O. sativa* and *O. glaberrima* described in our previous study(1) to identify insertions larger than 50 bp. Insertions were screened for homology with TE sequences by blastn against the TREP database (wheat.pw.usda.gov/ggpages/Repeats/). Using an in-house Perl script, TE with the highest homology were mapped onto the *O. sativa/O. glaberrima* alignments to facilitate visual inspection and to classify the polymorphism as transposon insertion or excision. Over 2,000 polymorphisms were screened, yielding the 482 insertions and 158 excisions (Supplementary Tables 1, 2 and 3).

Test for orthology of the analysed loci

To ensure that the aligned sequences from *O. sativa* and *O. glaberrima* indeed come from orthologous loci, we mapped the sequences used for the alignments back onto both genomes. That is, the sequences from *O. sativa* were first mapped back to the *O. sativa* genome and then mapped on to the *O. glaberrima* genome. The same was done *vice versa* with the corresponding *O. glaberrima* sequence (see methods). We split the aligned 24 kb regions into segments of 1000 bp and mapped each segment by blastn to the genome it came from as well as to the genome of the other species. This was done because blast alignments are often fragmented due to the presence of low-complexity sequences or TE insertions in one or the other species. Therefore, one can not expect a long sequence from one species to produce a similarly long blast hit in another. We therefore rather assigned each locus a score for how many of the segments map in the putative orthologous region in the other genome as a quantitative assessment of how strong the evidence for true orthology is for a particular locus.

For each 1000 bp segment, we recorded the positions of the top blast hit in the genome it came from as well as to the genome of the other species. We required that the top blast hit produced an alignment of at least 600 bp. Thus, some segments could not be mapped due to the presence of low-complexity sequences that are filtered out in the blastn search. Furthermore, one expects that not all segments map unambiguously to the orthologous locus in the other genome. This can, for example, be due to a large retrotransposon insertion in one species. The segments covering that retrotransposon would have no counterpart in the orthologous locus in the other species and therefore map elsewhere in the genome. The genomic region where the majority of the segments map was considered the putative ortholog. Furthermore, since we ran the analysis in both directions, we required that sequences from both species had to identify each other as the closest homolog. All analysed loci fulfilled these criteria. Additionally, as Supplementary Figure 2 shows, all except two loci are located in perfect colinear order along the chromosomes.

Classification of transposon polymorphisms into insertions and excisions

We defined a TE polymorphisms as an insertion if one species contained the TE plus the duplicated target site (TSD) on both sides, while the other species only contained one copy of the target site.

Excision are more difficult to define as they can go along with various re-arrangements (1,2). In general, we defined an excision by the absence of the TE in one species, with the pattern differing from that of an insertion. We distinguished different types of excisions: (i) in a perfect excision, as previously defined (2), one species contains the TE with flanked by the two units of the TSD wile the other species does not contain the TE but both copies of the TSD. (ii) Excisions with deletions were defined a the TE plus some flanking sequences being absent in one species. To distinguish these events from random deletions that by chance removed the TE plus flanking regions, we requested that one breakpoint of the excision be within 3 bp of one end of the TE (we considered it unlikely that a random deletion would have one of its borders so close to the end of a TE). (iii) Excisions with fillers were defined as events where the TE in one species is replaced with a completely unrelated sequence in the other. Fillers can range from a few bp to several kb. Also here, we requested that end of the filler sequence be within 3 bp of one end of the TE. Filler insertions were often found combined with deletions as described in (ii).

Quantification of mutations in sequences flanking TE polymorphisms

For all identified insertions and excisions, 12 kb of the flanking sequences were extracted from the *O. sativa* and *O. glaberrima* genome-wide alignment. We selected all alignments where more than 7,000 bases could be aligned (due to large insertions and deletions and/or colinearity breaks, usually less than 12 kb were actually aligned). This selection resulted in 206 sequence alignments for excisions and 438 for insertions. The transposon excision/insertion site was used as anchor point (i.e. position zero) from which all nucleotide substitutions and InDels were recorded. Sequence polymorphisms were added up for all alignments relative to the TE excision/insertion site. For the graphical representation (Figure 2A and B), nucleotide substitutions and InDel densities were calculated by a running average.

Comparison of promoter sequences from *O. sativa* and *O. glaberrima*

Information on start and end point of genes was extracted from the gff format annotation of the rice genome. As start and end point of genes we used transcription start and end points. Here, we used

rice genome version 5, because our previously published genome alignment of *O. sativa* and *O. glaberrima* (1) was done with this version. We defined the region from the transcription start point to 2 kb upstream of it as promoter region. Alignments were accepted when more than 600 bp in this 2 kb region could be aligned between *O. sativa* and *O. glaberrima*. For comparison, alignments of intergenic sequences were used. Here, we isolated segments that are located in the middle of intergenic sequence that are at least 10 kb in size (i.e. the distance between the end of one gene and the start of the next one is over 10 kb). Because sequence conservation along chromosome varies (Figure 3), chromosome arms were divided into 3 equally-sized bins for comparison of promoter and intergenic sequences. Data for promoters and intergenic sequences were analysed separately for each chromosome bin. To test whether the datasets for the individual bins differ from each other, the wilcox.test program from RStudio (rstudio.com) was used.

Comparison of CDS of genes

Repositories where coding sequences (CDS) of different species were obtained are listed in Supplementary Table 6. CDS for *O. glaberrima* were deduced from aliment with *O. sativa* CDS and are available upon request. Closest homologs from different species or, in the case of maize, homeologs that originated from a whole-genome duplication were identified by bi-directional blastn searches. Only homologs which had each other as the top blastn hit were used for comparison. Bi-directional closest homologs were aligned at the protein level using the program WATER from the EMBOSS package (emboss.sourceforge.net). The aligned protein sequences were back-translated to ensure that corresponding codons were aligned. We considered only alignment positions corresponding to the third codon base for Ala, Gly, Leu, Pro, Arg, Ser, Thr and Val. For those amino acids which all have six possible codons (Leu, Arg and Ser), we used only the codons starting with CT, TC and CG, respectively (i.e. the codons in which the third base can be exchanged without causing an amino acid change). To normalize the different sizes of genes, the aligned CDS were split into 5 equally-sized bins. To obtain sufficiently high numbers of synonymous substitutions, we used only gene pairs where more than 1,500 bp of the CDS could be aligned. For each bin of each gene, we calculated the number of synonymous substitutions per kb. Finally, we compiled the data for the

five bins for all genes. To test whether the datasets for the individual bins differ from each other, the wilcox.test program from RStudio was used.

De novo identification of small non-autonomous DNA transposons in Arabidopsis

DNA transposons are characterised by the presence of terminal inverted repeats (TIRs) which serve as binding site for transposase enzymes (3). The initial step of *de novo* identification was to screen chromosomal segments in windows of 1,000 bp, which overlap by 500 bp. The 1,000 windows were aligned with the program WATER from the EMBOSS package against themselves in reverse orientation. Outputs were parsed and visually inspected for the presence of inverted repeats longer than ~15 bp and over ~70% identity. The candidate sequences (inverted repeat and the sequences between them) were excised from the 1,000 bp. The candidate TEs were then used in blastn searches against the respective genome. Sequences with multiple hits were considered true DNA transposons. The *de novo* detection was done on one entire Arabidopsis chromosome, 2 Mbp of poplar linkage group 1 and 500 kb of rice chromosome 10 (Supplementary Figure 7).

Comparative analysis of DNA methylation

Data on methylation sites in *O. sativa* and *O. glaberrima* were kindly provided by Detlef Weigel and Claude Becker (Max Planck Institute for Developmental Biology, Tuebingen, Germany). These datasets will be published elsewhere. Sequence segments of 4 kb spanning the polymorphic transposon in *O. sativa* and *O. glaberrima* were extracted from the chromosomes. Methylated sites were flagged and the sequence segments were aligned with the program Water (emboss package, emboss.sourceforge.net/). Since we found that practically no methylation sites were conserved between the two species, methylation states were compared by simply counting the numbers of methylated sites in the sequences segments from the two species. The ratio of the number of methylation sites in *O. sativa* and *O. glaberrima* was then calculated for each transposon locus. For comparison, a second segment 2,000-4,000 bp downstream of the transposon was extracted.

Data access

Genome sequences used for the analyses are publicly available. Sequence aliments of genomic and CDS sequences are available upon request. Correspondence and requests for materials should be addressed to T.W. (wicker@botinst.uzh.ch).

References

1. Roffler S, Wicker T. 2015. Genome-wide comparison of Asian and African rice reveals high recent activity of DNA transposons. Mob DNA. 2015; **6:**8.

2. Buchmann JP, Matsumoto T, Stein N, Keller B, Wicker T. Interspecies sequence comparison of Brachypodium reveals how transposon activity corrodes genome colinearity. Plant J. 2012; **488:** 213-217.

3. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P and Schulman AH. A unified classification system for eukaryotic transposable elements. Nat Rev Genet 2007; **8:** 973-982.

Supplementary Note

Contents

A. Transposable elements and their contribution to evolution

DNA transposons can excise from the genome and re-insert elsewhere. When transposons excise, they leave double-strand breaks (DSBs) that have to be repaired by the cell. Depending on the repair pathway, this can lead to deletions and/or insertions of “filler” sequences at the site of the DSB (1-3). The initial step in DSB repair is usually the generation of 3' overhangs through exonucleases at the site of the break. Depending on the time that elapses before other repair enzymes are recruited, these 3' overhangs can be several kb in size, at least in yeast (4). The 3' overhangs can directly anneal to each other by single-strand annealing (SSA), using a few bp of micro-homology (reviewed by 5,6). This ultimately leads to a deletion of the segment between the annealing motifs. Previous studies showed that such deletions can range from a few bp (1,3) to several kb (2,3). Alternatively, a 3' overhang can invade a foreign DNA strand and use it as an intermediate template for DNA synthesis in a process called synthesis-dependent strand annealing (5-7). This leads to the introduction of a copy of the foreign template at the DSB site. Repair is completed when the leftover single-stranded DNA segments are used as templates for the synthesis of a new second strand. Sometimes, deletions and filler insertions at the excision site can be so extensive that transposon excisions are very difficult to identify as such, thereby explaining the generally low number of identified excisions

(2,3).

How much transposable elements (TEs) contribute to the evolution of genes and species is still unclear. Certainly, there have been cases where TEs contributed to major evolutionary innovations. For example the V(D)J recombination in the vertebrate immune system most likely has its origin in a transposable element (8). Additionally, there have been several studies showing that TEs can generate novel genic sequences, for example through gene retrotransposition or by providing new exons in a process called exonization (9). There are also many studies that described their influence on gene expression (example in 10). Thus, evidence for TE-driven evolutionary innovation is patchy and often anecdotal and the quantitative contribution of TEs to genome evolution is still unknown (9,10).

B. Background on grass comparative genomics

Grasses evolved from a common ancestor approximately 70 Myr ago (11). They are part of the major plant group of the monocotyledons which diverged from its "sister" group, the dicotyledons, approximately 145-300 Myr ago (12,13). Grasses provide an excellent dataset for comparative analyses because the genomes of representatives of the major clades *Bambusoideae*, *Ehrhartoideae* and *Pooideae* have been sequenced. This allows comparative analyses between clades, for example between the genomes of rice (14) and maize (15) as well as within clades, for example of wheat (16) and barley (17).

One unique characteristic of the rice and other grass genomes is that they contain enormous numbers of DNA (Class 2) transposons. Indeed, the superfamilies *DTT_Mariner* and *DTH_Harbinger* alone are present in at least 40,000 copies in grass genomes (15,18). Most DNA transposons described to date in grasses are small non-autonomous derivatives which do not encode any proteins and which depend for their transposition on transposase enzymes that are encoded by a small number of autonomous "mother" elements (18,19). Some of the non-autonomous elements (mostly those of the *DTT_Mariner* and *DTH_Harbinger* superfamilies) are referred to as miniature inverted-repeat transposable elements (20,21). Due to their small size they only contribute relatively little to the overall genome size and often seem to be tolerated in or near genes (18,20,21).

C. Methodological considerations on distinguishing transposon excisions from insertions

It is surprisingly difficult to identify transposon excision events in a comparative analysis. It was therefore essential to our study that we could distinguish transposon excisions from insertions with high confidence. We defined stringent criteria for an event to be classified as an excision, and preferred to discard unclear events. Previous studies showed that transposons excisions can produce a variety of patterns, including deletions and insertions of filler sequences (1,2,3,22). Since deletions and filler insertions can obscure excisions beyond recognition, or because deletions could by chance remove entire transposons, we required that at least one breakpoint of the deletion of filler insertion be within 3 bp of one end of the transposon (we considered it unlikely that a random deletion would have one of its borders so close to the end of a TE).

Furthermore, it is possible that some events we classified as insertions are in fact excisions that removed the transposon and precisely one copy of the target site. Such events were defined as "precise" excisions by Yang et al. (22). In a comparative analysis such as ours, it is impossible to distinguish precise excisions from insertion events. Interestingly, there are conflicting reports on the frequency of precise excisions. Using a heterologous system expressing the rice mPing element in Arabidposi, Yang et al. (22) reported that 25 of 30 excisions were precise. In contrast, Kikuchi et al. (23), working with the same element in rice anther cultures, found only one out of approximately 70 excision events to be precise. Also our own data suggest that the proportion of precise excisions may be small: we compared transposon polymorphisms which we classified as insertions with insertions of *Gypsy* retrotransposons (which can not excise). Both show similar increased mutation frequencies in their flanking regions, indicating that insertions also induce mutations in nearby sequences (which is not surprising, since the insertion process also has single-stranded intermediates). Nevertheless, insertions show overall much fewer mutations in their flanking regions than events that were classified as excisions (see Figure 3; Supplementary Figure 3). From this, we conclude that our criteria indeed distinguish different types of events (i.e. excisions and insertions) and that the events we classified as insertions contain only few precise excisions.

D. Test for orthology of compared sequence segments

Because we make a major claim about the role of TEs in evolution, it is important that concerns over potential weaknesses are addressed in detail. Thus, critical factors in our methods as well as in the interpretation of the results are discussed in the following. A crucial part of our case was to make sure that indeed orthologous loci were compared. Otherwise one could argue that putative excision sites that contain many polymorphisms are simply distant paralogs of which one never actually contained a transposon. Independent mapping of the analyzed sequences back onto the genomes showed that the analyzed loci all have exactly one homolog in each of the species, with almost all putative orthologs being located in colinear positions along chromosomes (Supplementary Fig 2). Theoretically, there is also the possibility that we compare deep paralogs, where a duplicated locus was present in the rice ancestor and subsequently, one copy was deleted in one species while the second copy was deleted in the other. This is a well-known problem in multi-copy gene families (example in 24). But sequence homology of such deep paralogs usually does not extend much past the sequences of the affected genes, while we aligned segments of up to 24 kb in size. We are thus confident that the vast majority of the sequences analyzed indeed represent orthologous loci.

E. Brassicaceae do not show increased mutation rates in termini of genes

To study whether the impact of DNA transposons is a general phenomenon in plants, we compared closest gene homologs in representatives of the dicotyledons which diverged from the monocotyledons about 145-300 Myr ago (12,13). We used multiple dicotyledon species, representing major lineages as well as different degrees of evolutionary distance. *Brassica rapa*, *B. napus*, *Arabidopsis thaliana* and *A. lyrata* were chosen as representatives of the *Brassicaceae* family. *A. thaliana* and *A. lyrata* diverged from each other approximately 10 Myr ago (see methods) while *Brassica* and *Arabidopsis* diverged approximately 32 Myr ago. Poplar (*Populus trichocarpa*) and soybean (*Glycine max*), which diverged approximately 70 Myr ago, were chosen as representatives of the *Fabid* clade. Interestingly, in none of the comparisons did we find increased substitution rates in terminal regions of genes (Figure 4D, Supplementary Figure 6, Supplementary Table 5), suggesting

that there is no effect of DNA transposons on genes comparable to that found in grasses.

Since we found a strong association of mutation rates in grass genes with DNA transposons activity, we expected that the genomes of dicotyledons contain fewer such elements. Therefore, we performed a *de novo* search for DNA transposons in the *A. thaliana* genome (see methods), in order to assess the abundance of these elements. Interestingly, we found only 27 different types of putative transposons, which were present in a total of 330 copies in *A. thaliana*. Furthermore, many of these elements are only fragments, as we classified only 65 as potentially intact elements. Thus, *A. thaliana* contains several orders of magnitude fewer DNA transposons than the grass genomes sequenced so far [8,28]. We also performed the *de novo* search in the *P. trichocarpa* genome which is with 495 Mbp even larger than the *O. sativa* genome. Here, we manually examined all 31 candidate transposons that were identified in the first 2 Mbp of linkage group 1. Only two turned out to be DNA transposons that are present at moderately high copy numbers (approximately 450 and 600 copies, respectively). In contrast, the same *de novo* search in only 500 kb in rice yielded 53 candidates, of which 20 had over 500 copies in the genome (Supplementary Figure 7).

F. Evaluation of evidence for transposons as the cause for increased mutation rates

Obviously, there are other possible causes for DSBs near genes besides transposon excisions, such as toxic chemicals, radiation or template breakage or slippage during replication. Following the repair pathway described in Figure 3, this could also lead to mutations during DSB repair. However, several lines of evidence support our claim that DNA transposons are at least a major factor leading to the elevated mutation rates in CDS and regulatory regions in grasses. First, our data from sequence comparisons show empirically that sequences flanking excisions contain highly elevated numbers of nucleotide substitutions and InDels. Since DNA transposons are strongly enriched in promoter and downstream regions, it follows that these regions will be disproportionately affected. We indeed find that promoters are one average less conserved than randomly picked intergenic sequences. Second, genes from *O. sativa* and *O. glaberrima* which have the highest sequence conservation, reflecting the overall genome-wide average, do not show a substitution rate gradient. In contrast, genes that have a below average sequence conservation show the gradient. Third, genomes which contain many DNA

transposons (such as grasses) all show the substitution rate gradient in genes, while those of dicotyledons (which contain much fewer DNA transposons) do not.

G. Comparative analysis of methylation states in polymorphic transposon loci

To study whether transposon excisions and insertions have an effect on the methylation state of the respective locus, we compared methylation data from *O. sativa* and *O. glaberrima* (see methods). Sequence segments of 4 kb spanning the polymorphic transposon in *O. sativa* and *O. glaberrima* were extracted from the chromosomes. The sequences were the aligned and positions of methylated bases compared. We we found that practically no methylation sites were conserved between the two species. Thus, overall methylation states were compared by simply counting the numbers of methylated sites in the sequences segments from the two species. The ratio of the number of methylation sites in *O. sativa* and *O. glaberrima* was then calculated for each transposon locus. For comparison, a second segment 2,000-4,000 bp downstream of the transposon was extracted. For excisions, we found a weak but significant (Wicoxon test p-value = 3.893e-05) difference in the two distributions (Supplementary Fig. 8). These data suggest that transposon excisions tend to be followed by de-methylation of the locus. For insertions the effect was weaker but still statistically significant (Wicoxon test p-value =0.008, Supplementary Fig. 8b). However, since practically no methylated sites were conserved in the two species and the loci studied, the described quantitative analysis is crude and we do not want to over-interpret these results.

References

1. Yang G., Weil CF. and Wessler SR. A rice Tc1/mariner-like element transposes in yeast. Plant Cell. 2006; **18:** 2469–2478.

2. Buchmann JP, Matsumoto T, Stein N, Keller B, Wicker T. Interspecies sequence comparison of Brachypodium reveals how transposon activity corrodes genome colinearity. Plant J. 2012; **488:** 213-217.

3. Roffler S, Wicker T. 2015. Genome-wide comparison of Asian and African rice reveals high recent activity of DNA transposons. Mob DNA. 2015; **6:**8.

4. Storici F, Snipe JR, Chan GK., Gordenin, D.A. Resnick, M.A. Conservative repair of a chromosomal double-strand break by single-strand DNA through two steps of annealing. J Cell Biol. 2006; **26:** 7645-7657.

5. Puchta, H. The repair of double-strand breaks in plants: Mechanisms and consequences for genome evolution. J Exp Botany. 2005; **56:** 1–14.

6. Hartlerode AJ, Scully R. Mechanisms of double-strand break repair in somatic mammalian cells. Biochem J. 2009; **423:** 157–168.

7. Nassif N, Penney J, Pal S, Engels WR, Gloor GB. Efficient copying of nonhomologous sequences from ectopic sites via P-element-induced gap repair. Mol Cell Biol 1994; **14:** 1613–1625.

8. Fugmann SD, Lee AI, Shockett PE, Villey IJ, Schatz DG. The RAG proteins and V(D)J recombination: complexes, ends, and transposition. Ann Rev Immunol. 2000; **18:** 495-527.

9. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. Nat Rev Genet. 2009; **10:** 691-703.

10. de Souza, F.S., Franchini, L.F. Rubinstein, M. Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong? Mol Biol Evol. 2013; **30:** 1239-1251.

11. Grass Phylogeny Working Group. Phylogeny and subfamilial classification of the grasses (Poaceae). Ann Mo Bot Gard. 2001; **88:** 373–457.

12. Kawai Y, Otsuka J. The deep phylogeny of land plants inferred from a full analysis of nucleotide base changes in terms of mutation and selection. J Mol Evol. 2004; **58:** 479–489.

13. Zimmer, A, Lang, D, Richardt, S, Frank, W, Reski, R and Rensing, SA. Dating the early evolution of plants: detection and molecular clock analyses of orthologs. Mol Gen Genomics. 2007; **278:** 393–402.

14. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. Nature. 2005; **436:** 793–800.

15. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F. et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 2009; **326:** 1112-1115.

16. International Wheat Genome Sequencing Consortium (IWGSC). A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. Science. 2014; **345:** 1251.

17. International Barley Genome Sequencing Consortium (IBSC), Mayer KF, Waugh R, Brown JW, Schulman AH et al. A physical, genetic and functional sequence assembly of the barley genome. Nature. 2012; **491:** 711-716.

18. International *Brachypodium* Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. Nature. 2010; **463:** 763–768.

19. Yang G., Weil CF. and Wessler SR. A rice Tc1/mariner-like element transposes in yeast. Plant Cell. 2006; **18:** 2469–2478.

20. Bureau T, Wessler SR. Mobile inverted-repeat elements of the Tourist family are associated with the genes of many cereal grasses. Proc Natl Acad Sci USA 1994; **9:** 907-916.

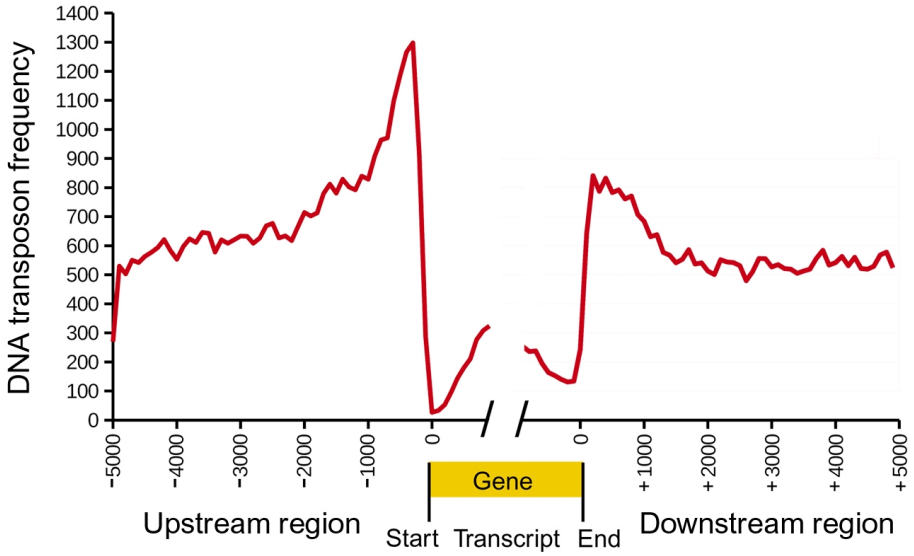
21. Bureau T, Wessler SR. Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Proc Natl Acad Sci USA 1994; **9:** 1411-1115.

22. Yang G, Zhang F, Hancock CN, Wessler SR. Transposition of the rice miniature inverted repeat transposable element mPing in Arabidopsis thaliana. Proc Natl Acad Sci USA. 2007; 104:10962-10967.

23 .Kikuchi K, Terauchi K, Wada M, Hirano HY. The plant MITE mPing is mobilized in anther culture. Nature. 2003; 421:167-70.

24. Bossolini E, Wicker T, Knobel PA, Keller B. Comparison of orthologous loci from small grass genomes Brachypodium and rice: implications for wheat genomics and grass genome annotation. Plant J. 2007; 49:704-717.

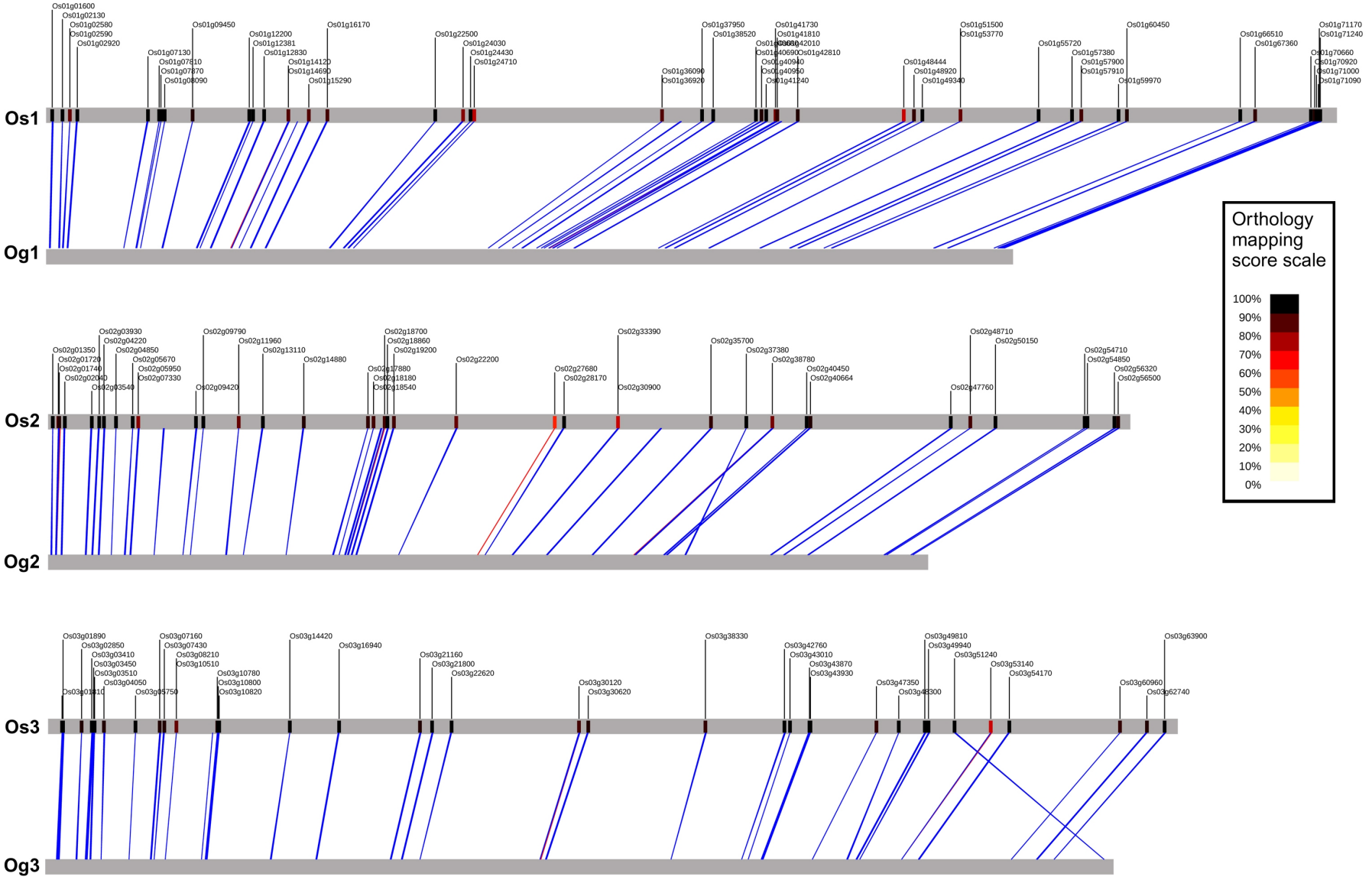
Supplementary Figures and Tables

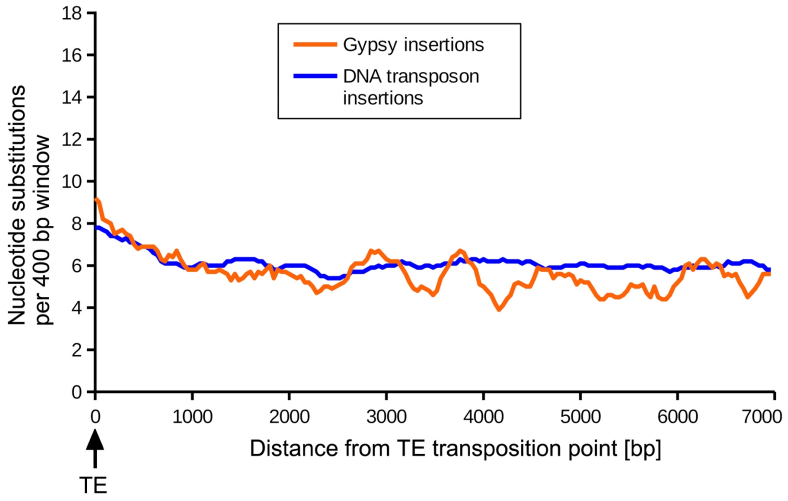


Supplementary Figure 1. Frequency of *DTT_Mariner* and *DTH_Harbinger* transposons relative to genes in *O. sativa*. Transposons in the up- and downstream regions of 21,444 genes were annotated and the cumulative occurrence plotted relative genes (e.g. the highest peak indicates that over 1,300 genes have a DNA transposon upstream of the transcription start point). The gene is shown in the center with 5,000 bp of up- and downstream region. Here, only genes longer than 2 kb were used. Thus, the center of the plot depicts the transposon frequencies of the 5' and 3' terminal 1,000 bp inside the genes.

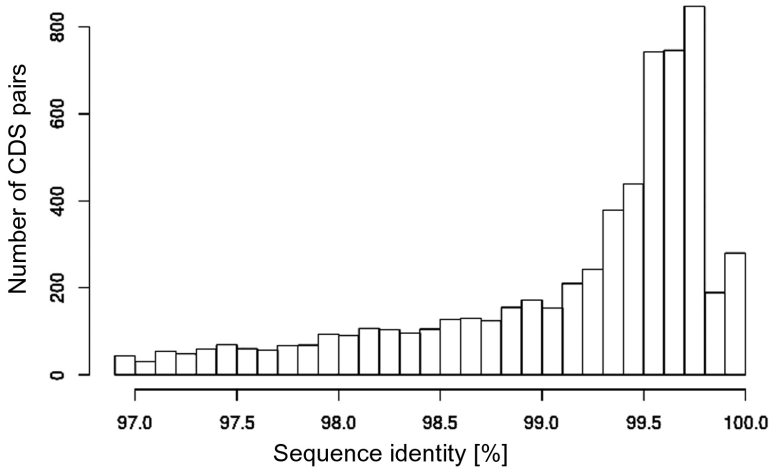
Following page:

Supplementary Figure 2. Test for orthology for the loci containing putative transposon excisions. For this study, we manually identified 158 loci from rice chromosomes 1, 2 and 3 containing putative excisions of DNA transposons in either *O. sativa* or *O. glaberrima*. The Figure shows the positions of the compared loci on the *O. sativa* (Os) and *O. glaberrima* (Og) chromosomes. Putative orthologous loci are connected with blue lines. Loci are named after the gene closest to the polymorphic transposon. Since we aligned up to 24 kb of the putative orthologous loci, segments of 1 kb were used to map the genomic sequences back to the genomes (see methods). Each locus was assigned a score describing the percentage of 1 kb segments that mapped to its putative ortholog counterpart in the other species (orthology mapping score). The score is indicated as a small vertical box in the *O. sativa* chromosome. Obviously, some pf the 1 kb segments may map elsewhere in the genome because they are comprised of polymorphic TEs or repetitive sequences that can not be mapped unambiguously. However, most loci have very high scores, indicating that most parts of the 24 kb sequences of one species map unambiguously to the putative orthologous locus in the other species. Furthermore, all expect 2 loci are located in perfect colinear order along the chromosomes (see also methods).

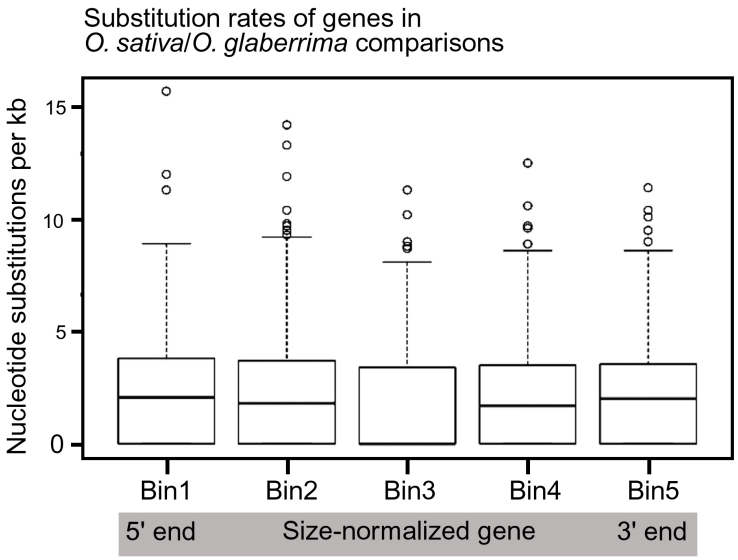

Os1
Og1
Os2
Og2
Os3
Og3
Orthology mapping score scale
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%



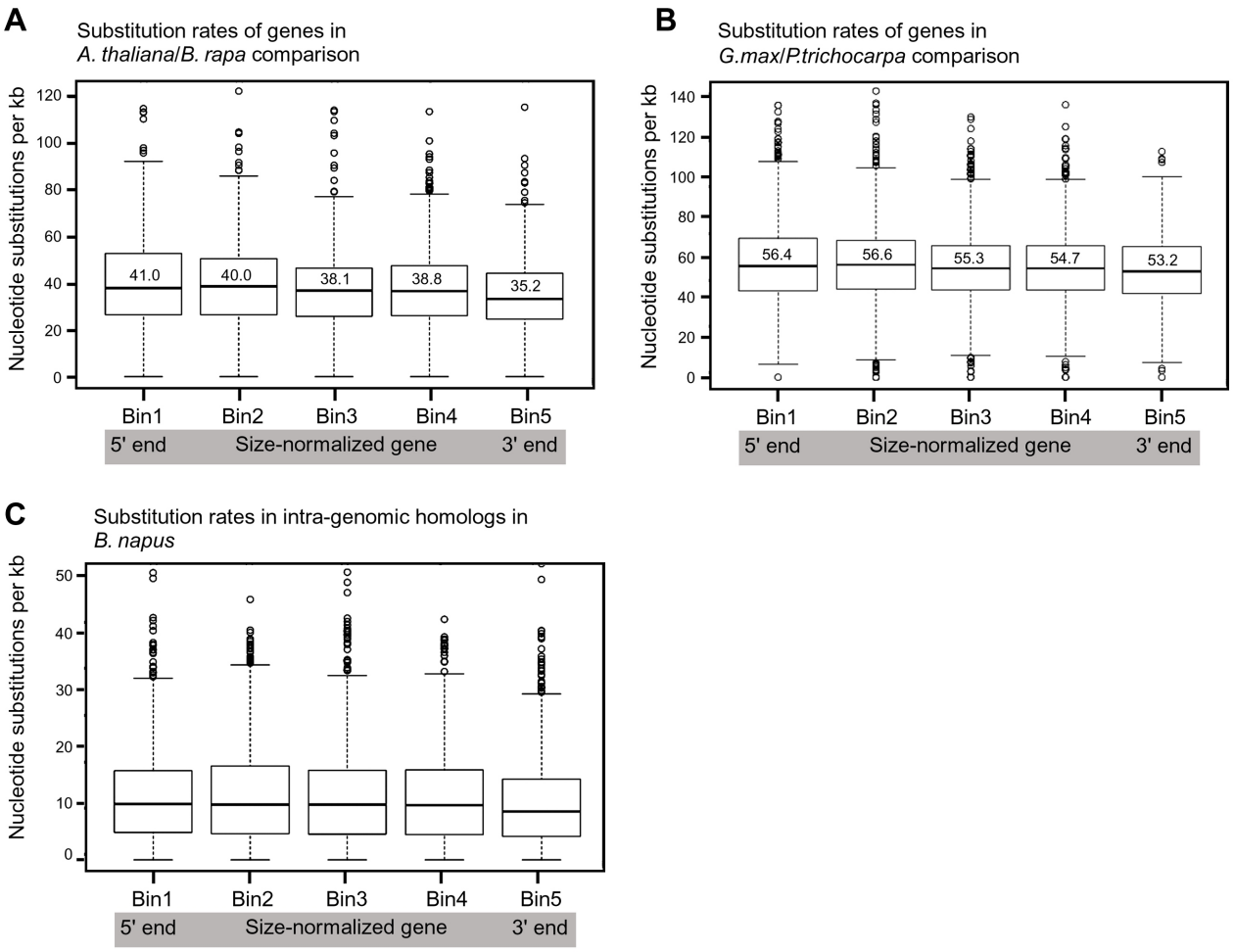
Supplementary Figure 3. Frequencies of nucleotide substitutions (NS) in relation to transposable elements insertion sites in rice. The plot shows a comparison of NS frequencies near insertions of DNA transposons (blue line) and Gypsy retrotransposons (red line). In both cases, NS frequency increases slightly toward the insertion point. This indicates that insertions also cause small numbers of mutations in their flanking sequences. Furthermore, this result is evidence that events classified as DNA transposon insertions probably do not contain many precise excisions. TE: transposable element insertion site.



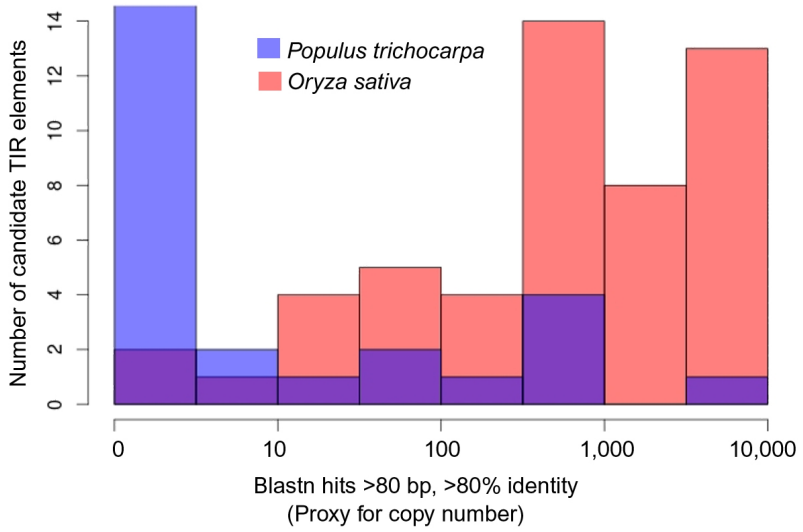
Supplementary Figure 4. Distribution of sequence identities of coding sequences (CDS) of closest homologs from *O. sativa* and *O. glaberrima*.



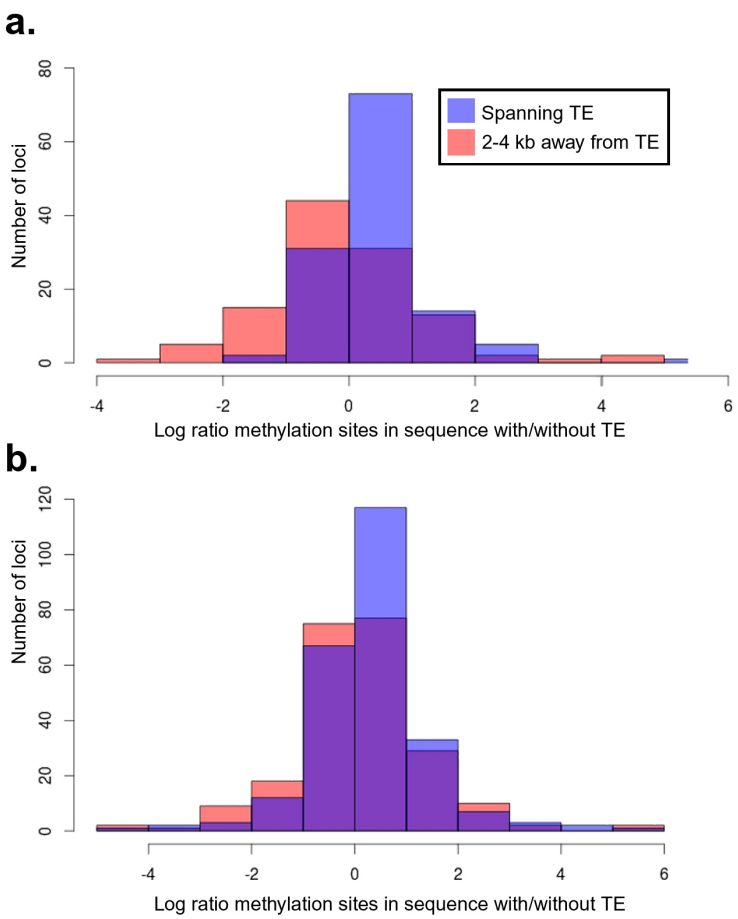
Supplementary Figure 5. Comparison of CDS of 312 genes from *O. sativa* and *O. glaberrima*. Here, only genes that are >99.5% identical (i.e. the overall level of sequence identity of the two genomes) were considered. The high conservation of these genes indicates that they were not affected by nearby error-prone DSB repair. They also do not show significantly lower sequence conservation in the center part of the gene.



Supplementary Figure 6. Nucleotide substitution frequencies in synonymous sites of genes. To normalize the different CDS sizes, genes were divided into 5 equally sized bins and frequencies were normalized to nucleotide substitutions per kb for each bin. The bold line inside the box is the median value, while mean values are indicated with numbers. (**A**) Comparison of 636 pairs of closest homologs from *A. thaliana* and *B. rapa*. (**B**) Comparison of 1,799 pairs of closest homologs from soybean (*G. max*) and poplar (*P. trichocarpa*). (**C**) Nucleotide substitution frequencies in synonymous sites of of 1,395 pairs of intra-genomic closest homologs in *B. napus* that originate from a whole genome duplication.



Supplementary Figure 7. Copy number estimates for candidate DNA transposons identified in *de novo* searches in the genomes of poplar (*P. trichocarpa*) and rice (*O. sativa*). As a proxy for copy numbers, each identified transposon candidates was used as a query in a blast search against its respective genome. All blast hits that were longer than 80 bp and >80% identical were considered. The x-axis shows the number of blast hits in a logarithmic scale while the y-axis shows the number of transposon candidates in each copy number range. Note that the *de novo* search in rice yielded many more elements which have on average much higher copy numbers than those in poplar.



Supplementary Figure 8. Comparative analysis of methylation data in loci containing polymorphic transposons. Numbers of methylation sites were compared in orthologous loci with and without transposons in *O. sativa* and *O. glaberrima.* For each locus, the ratio of the numbers of methylated sites was calculated. The figure shows the distribution of the Log10 of these ratios. To study the effect of transposon insertions and excisions, data from 4 kb segments spanning the transposon (blue) site were compared with data from segments covering the sequence 2,000-4,000 bp away from the transposon (red). **a.** Datasets for transposon excisions. **b.** Datasets for transposon insertions. Note that in both datasets the ratio of numbers for sequence with transposon/sequence without transposon are shifted towards higher values, indicating that sequence segments containing transposons tend to have more methylated sites.

Supplementary Table 1. Positions of DNA (Class 2) transposon excisions in the two rice species *O. sativa* and *O. glaberrima*. Chromsomal positions are given for *O. sativa* genome version6 and *O. glaberrima* genome version 1. OsChr: *O. sativa* chromosome. OsPos: base pair position on *O. sativa* chromsome. OgChr: *O. glaberrima* chromosome. OgPos: base pair position on *O. glaberrima* chromsome.

OsChr	OsPos	OgChr	OgPos	Event
1	306959	1	211329	excision in *O. glaberrima*
1	616788	1	508545	excision in *O. sativa*
1	858433	1	627119	excision in *O. glaberrima*
1	864565	1	637088	excision in *O. glaberrima*
1	1074814	1	768963	excision in *O. glaberrima*
1	3360667	1	2590082	excision in *O. glaberrima*
1	3745430	1	2979321	excision in *O. sativa*
1	3803142	1	3006536	excision in *O. glaberrima*
1	3918145	1	3132783	excision in *O. glaberrima*
1	4815386	1	3847255	excision in *O. glaberrima*
1	6651580	1	4959979	excision in *O. sativa*
1	6765422	1	5047554	excision in *O. glaberrima*
1	7102225	1	5394206	excision in *O. sativa*
1	7925022	1	6070534	excision in *O. glaberrima*
1	8213940	1	6325790	excision in *O. sativa*
1	8553288	1	6685238	excision in *O. glaberrima*
1	9141251	1	7174597	excision in *O. sativa*
1	12649204	1	9251957	excision in *O. glaberrima*
1	13541075	1	9696587	excision in *O. sativa*
1	13772287	1	9835114	excision in *O. glaberrima*
1	13903805	1	10013470	excision in *O. sativa*
1	19973277	1	14353253	excision in *O. glaberrima*
1	20578900	1	14685479	excision in *O. sativa*
1	21259318	1	15128202	excision in *O. sativa*
1	21634661	1	15450764	excision in *O. glaberrima*
1	22993368	1	15917686	excision in *O. sativa*
1	22993368	1	15917686	excision in *O. sativa*
1	23164571	1	16062445	excision in *O. glaberrima*
1	23165619	1	16063174	excision in *O. glaberrima*
1	23342045	1	16158338	excision in *O. sativa*
1	23625935	1	16303696	excision in *O. glaberrima*
1	23675316	1	16441712	excision in *O. glaberrima*
1	23814701	1	16579729	excision in *O. sativa*
1	24364456	1	17122959	excision in *O. sativa*
1	27781873	1	19847466	excision in *O. glaberrima*
1	28072592	1	20070363	excision in *O. glaberrima*
1	28361571	1	20364679	excision in *O. sativa*
1	29591258	1	21479821	excision in *O. sativa*
1	29591258	1	21479821	excision in *O. sativa*
1	32090182	1	23123244	excision in *O. glaberrima*
1	33174763	1	24094927	excision in *O. sativa*
1	33480796	1	24350471	excision in *O. glaberrima*
1	33483408	1	24353462	excision in *O. sativa*
1	34672522	1	25183221	excision in *O. sativa*
1	34960832	1	25434965	excision in *O. sativa*
1	38626114	1	28741511	excision in *O. glaberrima*
1	39098738	1	29191544	excision in *O. glaberrima*
1	40901397	1	30670845	excision in *O. glaberrima*
1	41051298	1	30810778	excision in *O. sativa*
1	41093097	1	30853689	excision in *O. glaberrima*
1	41144155	1	30907102	excision in *O. glaberrima*
1	41177892	1	30952903	excision in *O. glaberrima*
1	41224131	1	31001343	excision in *O. glaberrima*
2	193005	2	144202	excision in *O. glaberrima*
2	400988	2	298559	excision in *O. sativa*
2	409744	2	310044	excision in *O. glaberrima*
2	570546	2	470835	excision in *O. sativa*

2	1441950	2	1258846	excision in *O. glaberrima*
2	1672674	2	1470629	excision in *O. glaberrima*
2	1852655	2	1673725	excision in *O. glaberrima*
2	2240965	2	2086762	excision in *O. sativa*
2	2774912	2	2521522	excision in *O. sativa*
2	2964408	2	2707426	excision in *O. glaberrima*
2	3778456	2	3461335	excision in *O. glaberrima*
2	4834349	2	4395923	excision in *O. glaberrima*
2	5051683	2	4634602	excision in *O. sativa*
2	6198543	2	5782266	excision in *O. sativa*
2	6963376	2	6351729	excision in *O. glaberrima*
2	8286770	2	7722270	excision in *O. sativa*
2	10361363	2	9244192	excision in *O. sativa*
2	10559040	2	9443107	excision in *O. sativa*
2	10805470	2	9624124	excision in *O. sativa*
2	10910114	2	9718146	excision in *O. glaberrima*
2	11008980	2	9839150	excision in *O. glaberrima*
2	11187934	2	9984296	excision in *O. glaberrima*
2	13233893	2	11353690	excision in *O. sativa*
2	16397439	2	13792955	excision in *O. sativa*
2	16672514	2	14149188	excision in *O. sativa*
2	18443613	2	15041579	excision in *O. sativa*
2	19833517	2	16144175	excision in *O. sativa*
2	21452668	2	17601177	excision in *O. sativa*
2	22574468	2	20611734	excision in *O. glaberrima*
2	23428543	2	18987091	excision in *O. sativa*
2	24523036	2	19921077	excision in *O. glaberrima*
2	24639439	2	20011778	excision in *O. glaberrima*
2	29194338	2	23380770	excision in *O. sativa*
2	29808797	2	23788631	excision in *O. glaberrima*
2	30632606	2	24572989	excision in *O. glaberrima*
2	33493057	2	27022104	excision in *O. sativa*
2	33582666	2	27103557	excision in *O. glaberrima*
2	34472769	2	27883966	excision in *O. sativa*
2	34589534	2	27943921	excision in *O. sativa*
3	499902	3	338565	excision in *O. sativa*
3	547017	3	384382	excision in *O. glaberrima*
3	1123914	3	964806	excision in *O. sativa*
3	1454688	3	1250217	excision in *O. glaberrima*
3	1491104	3	1284691	excision in *O. glaberrima*
3	1520100	3	1403811	excision in *O. glaberrima*
3	1858065	3	1754043	excision in *O. glaberrima*
3	2863754	3	2654354	excision in *O. glaberrima*
3	3661860	3	3356319	excision in *O. glaberrima*
3	3770950	3	3462134	excision in *O. glaberrima*
3	4185134	3	3828737	excision in *O. glaberrima*
3	5351516	3	5000021	excision in *O. glaberrima*
3	5505760	3	5126520	excision in *O. glaberrima*
3	5536148	3	5155184	excision in *O. glaberrima*
3	5551093	3	5170411	excision in *O. sativa*
3	7843648	3	7233157	excision in *O. glaberrima*
3	9419504	3	8701806	excision in *O. glaberrima*
3	12064524	3	11110004	excision in *O. sativa*
3	12455139	3	11463171	excision in *O. sativa*
3	13039003	3	12046638	excision in *O. glaberrima*
3	17195722	3	15931687	excision in *O. glaberrima*
3	17466945	3	16114828	excision in *O. sativa*
3	21272026	3	20148148	excision in *O. sativa*
3	23810131	3	22424470	excision in *O. glaberrima*
3	23988359	3	22633235	excision in *O. sativa*
3	24606514	3	23056031	excision in *O. sativa*
3	24647524	3	23102948	excision in *O. glaberrima*
3	26773320	3	24710738	excision in *O. glaberrima*
3	27496942	3	25840716	excision in *O. sativa*
3	28358918	3	26130190	excision in *O. sativa*
3	28466812	3	26233042	excision in *O. sativa*

3	29307053	3	34117698	excision in *O. glaberrima*
3	30468998	3	27578832	excision in *O. glaberrima*
3	31048314	3	28131446	excision in *O. sativa*
3	34631324	3	31127814	excision in *O. glaberrima*
3	35491552	3	31944763	excision in *O. sativa*
3	36097960	3	32506454	excision in *O. sativa*

Supplementary Table 2. Positions of DNA (Class 2) transposon insertions in the two rice species *O. sativa* and *O. glaberrima*. Chromsomal positions are given for *O. sativa* genome version6 and *O. glaberrima* genome version 1. OsChr: *O. sativa* chromosome. OsPos: base pair position on *O. sativa* chromsome. OgChr: *O. glaberrima* chromosome. OgPos: base pair position on *O. glaberrima* chromsome.

OsChr	OsPos	OgChr	OgPos	Event
2	548163	2	431986	insertion in *O. sativa*
2	1007456	2	841708	insertion in *O. glaberrima*
2	1097229	2	939142	insertion in *O. sativa*
2	1394657	2	1210000	insertion in *O. glaberrima*
2	1451345	2	1267611	insertion in *O. sativa*
2	1521432	2	1322748	insertion in *O. sativa*
2	3229543	2	2909625	insertion in *O. sativa*
2	3301024	2	2981497	insertion in *O. glaberrima*
2	3653146	2	3337718	insertion in *O. sativa*
2	3674570	2	3359044	insertion in *O. sativa*
2	3757652	2	3442306	insertion in *O. glaberrima*
2	3928217	2	3565077	insertion in *O. sativa*
2	4147768	2	3688881	insertion in *O. sativa*
2	4290583	2	3838942	insertion in *O. sativa*
2	4486929	2	4023272	insertion in *O. sativa*
2	4622678	2	4166409	insertion in *O. sativa*
2	4654407	2	4201648	insertion in *O. glaberrima*
2	4752142	2	4317340	insertion in *O. glaberrima*
2	5190270	2	4767482	insertion in *O. sativa*
2	5235725	2	4812481	insertion in *O. sativa*
2	5657121	2	5243037	insertion in *O. glaberrima*
2	5855629	2	5442276	insertion in *O. glaberrima*
2	5906514	2	5503460	insertion in *O. glaberrima*
2	5955309	2	5569589	insertion in *O. glaberrima*
2	6252471	2	5839700	insertion in *O. glaberrima*
2	6262767	2	5851858	insertion in *O. glaberrima*
2	6431234	2	6037740	insertion in *O. sativa*
2	6783920	2	6161665	insertion in *O. glaberrima*
2	6814392	2	6193919	insertion in *O. sativa*
2	6906056	2	6270107	insertion in *O. sativa*
2	7013533	2	6401988	insertion in *O. sativa*
2	7136689	2	6548048	insertion in *O. sativa*
2	7302742	2	6707731	insertion in *O. sativa*
2	7760791	2	7165360	insertion in *O. glaberrima*
2	8992287	2	8328170	insertion in *O. sativa*
2	9070594	2	8405295	insertion in *O. glaberrima*
2	9113359	2	8429490	insertion in *O. sativa*
2	9410673	2	8553323	insertion in *O. sativa*
2	10058277	2	9122738	insertion in *O. sativa*
2	10533680	2	9417561	insertion in *O. sativa*
2	10720163	2	9538327	insertion in *O. sativa*
2	10779807	2	9596088	insertion in *O. sativa*
2	10959876	2	9789025	insertion in *O. glaberrima*
2	11061761	2	9884599	insertion in *O. sativa*
2	14751202	2	12343679	insertion in *O. sativa*
2	15674238	2	13441149	insertion in *O. glaberrima*
2	16210731	2	13738559	insertion in *O. sativa*
2	17196696	2	14503969	insertion in *O. glaberrima*
2	17225368	2	14532168	insertion in *O. glaberrima*
2	18690505	2	15213556	insertion in *O. sativa*
2	18816844	2	15326737	insertion in *O. glaberrima*
2	19205531	2	15599243	insertion in *O. sativa*
2	19564156	2	15918400	insertion in *O. sativa*
2	19584120	2	15936738	insertion in *O. sativa*
2	19740850	2	16053640	insertion in *O. sativa*
2	19958660	2	16274099	insertion in *O. sativa*
2	20124694	2	16443706	insertion in *O. sativa*

2	20157721	2	16478652	insertion in *O. glaberrima*
2	20946541	2	17154650	insertion in *O. sativa*
2	21082955	2	17299069	insertion in *O. sativa*
2	21543303	2	17683708	insertion in *O. sativa*
2	21563217	2	17702980	insertion in *O. glaberrima*
2	21921457	2	17983530	insertion in *O. sativa*
2	22070941	2	18147353	insertion in *O. sativa*
2	22406944	2	18288259	insertion in *O. sativa*
2	22495996	2	22495996	insertion in *O. glaberrima*
2	22541034	2	20575505	insertion in *O. sativa*
2	22554806	2	20587784	insertion in *O. glaberrima*
2	22642951	2	18374193	insertion in *O. sativa*
2	22804060	2	18458294	insertion in *O. sativa*
2	22815346	2	18479046	insertion in *O. sativa*
2	23448710	2	18987738	insertion in *O. glaberrima*
2	23473847	2	19003942	insertion in *O. sativa*
2	23824527	2	19327044	insertion in *O. sativa*
2	23974169	2	19453644	insertion in *O. glaberrima*
2	24131443	2	19602959	insertion in *O. sativa*
2	24235675	2	19707384	insertion in *O. sativa*
2	24268744	2	19735158	insertion in *O. sativa*
2	24470903	2	19868817	insertion in *O. glaberrima*
2	24677182	2	20045537	insertion in *O. sativa*
2	25045807	2	20393831	insertion in *O. sativa*
2	25152424	2	20476874	insertion in *O. sativa*
2	25810130	2	20754995	insertion in *O. sativa*
2	26518050	2	21391569	insertion in *O. sativa*
2	27157357	2	21913411	insertion in *O. glaberrima*
2	27499344	2	22232412	insertion in *O. sativa*
2	27731114	2	22307406	insertion in *O. sativa*
2	28006147	2	22519911	insertion in *O. sativa*
2	28086917	2	22574545	insertion in *O. sativa*
2	28260263	2	22718849	insertion in *O. glaberrima*
2	28461725	2	22906233	insertion in *O. glaberrima*
2	28489288	2	22933156	insertion in *O. sativa*
2	29022471	2	23268211	insertion in *O. glaberrima*
2	29539680	2	23518558	insertion in *O. sativa*
2	29714180	2	23694148	insertion in *O. sativa*
2	29853203	2	23836420	insertion in *O. glaberrima*
2	30272346	2	24223963	insertion in *O. sativa*
2	30852647	2	24772078	insertion in *O. glaberrima*
2	32236122	2	25945231	insertion in *O. sativa*
2	32345377	2	26059054	insertion in *O. sativa*
2	32631080	2	26302291	insertion in *O. sativa*
2	32677692	2	26350340	insertion in *O. glaberrima*
2	32736887	2	26411599	insertion in *O. sativa*
2	32884148	2	26522127	insertion in *O. sativa*
2	33403287	2	26932179	insertion in *O. glaberrima*
2	33929106	2	27446500	insertion in *O. sativa*
2	34045288	2	27552982	insertion in *O. sativa*
2	34069760	2	27566980	insertion in *O. glaberrima*
2	34760092	2	28091069	insertion in *O. sativa*
2	34784954	2	28107797	insertion in *O. sativa*
2	34812529	2	28124520	insertion in *O. sativa*
2	35049936	2	28346461	insertion in *O. sativa*
2	35094170	2	28387094	insertion in *O. sativa*
2	35228196	2	28522776	insertion in *O. glaberrima*
2	35254201	2	28548082	insertion in *O. sativa*
2	35499374	2	28755864	insertion in *O. glaberrima*
2	35775697	2	28976705	insertion in *O. glaberrima*
3	179673	3	70226	insertion in *O. glaberrima*
3	432251	3	258020	insertion in *O. sativa*
3	483767	3	320333	insertion in *O. glaberrima*
3	603926	3	452443	insertion in *O. glaberrima*
3	742408	3	613190	insertion in *O. glaberrima*
3	1102395	3	944087	insertion in *O. sativa*

3	1241821	3	1074320	insertion in *O. sativa*
3	1277136	3	1097455	insertion in *O. sativa*
3	1300289	3	1117979	insertion in *O. sativa*
3	1532872	3	1416234	insertion in *O. sativa*
3	1784261	3	1666803	insertion in *O. sativa*
3	2229794	3	2097495	insertion in *O. sativa*
3	2419617	3	2241107	insertion in *O. sativa*
3	2459513	3	2275095	insertion in *O. sativa*
3	2749266	3	2545522	insertion in *O. sativa*
3	2810787	3	2602425	insertion in *O. glaberrima*
3	3177063	3	2921689	insertion in *O. glaberrima*
3	3272423	3	3009361	insertion in *O. glaberrima*
3	3493104	3	3209778	insertion in *O. sativa*
3	3753871	3	3445721	insertion in *O. sativa*
3	3792915	3	3481968	insertion in *O. glaberrima*
3	3831566	3	3520761	insertion in *O. sativa*
3	3967050	3	3618879	insertion in *O. sativa*
3	4136862	3	3784486	insertion in *O. sativa*
3	4853429	3	4488711	insertion in *O. sativa*
3	4907805	3	4548231	insertion in *O. sativa*
3	4945022	3	4584987	insertion in *O. sativa*
3	5632193	3	5250175	insertion in *O. sativa*
3	5655113	3	5271844	insertion in *O. glaberrima*
3	5714478	3	5335157	insertion in *O. sativa*
3	6153041	3	5776098	insertion in *O. sativa*
3	6951224	3	6390831	insertion in *O. sativa*
3	7006606	3	6450862	insertion in *O. glaberrima*
3	7093423	3	6537106	insertion in *O. glaberrima*
3	7194764	3	6615449	insertion in *O. sativa*
3	7229112	3	6649341	insertion in *O. glaberrima*
3	7594402	3	6991535	insertion in *O. glaberrima*
3	8097631	3	7487581	insertion in *O. sativa*
3	8350267	3	7723018	insertion in *O. glaberrima*
3	8506624	3	7861374	insertion in *O. sativa*
3	8597087	3	7940642	insertion in *O. glaberrima*
3	9013759	3	8330236	insertion in *O. sativa*
3	9154374	3	8450855	insertion in *O. sativa*
3	9216918	3	8500219	insertion in *O. glaberrima*
3	9356255	3	8640143	insertion in *O. glaberrima*
3	9505157	3	8786476	insertion in *O. sativa*
3	9739063	3	9012789	insertion in *O. sativa*
3	9964979	3	9241000	insertion in *O. sativa*
3	10107149	3	9298790	insertion in *O. sativa*
3	10234031	3	9421802	insertion in *O. sativa*
3	10419377	3	9611882	insertion in *O. sativa*
3	10945242	3	10103884	insertion in *O. glaberrima*
3	11254755	3	10315225	insertion in *O. sativa*
3	11480938	3	10512763	insertion in *O. glaberrima*
3	11512629	3	10555574	insertion in *O. sativa*
3	11583782	3	10631534	insertion in *O. glaberrima*
3	11721018	3	10778444	insertion in *O. sativa*
3	12737874	3	11759659	insertion in *O. sativa*
3	12760204	3	11774293	insertion in *O. sativa*
3	12805314	3	11811283	insertion in *O. sativa*
3	12938738	7	25264645	insertion in *O. sativa*
3	12960087	7	25421746	insertion in *O. glaberrima*
3	13005184	3	12012637	insertion in *O. sativa*
3	13260643	3	12272860	insertion in *O. sativa*
3	13700648	3	12572240	insertion in *O. sativa*
3	13712575	3	12584377	insertion in *O. glaberrima*
3	13829336	3	12706605	insertion in *O. sativa*
3	13884561	3	12746803	insertion in *O. sativa*
3	14292092	3	13096548	insertion in *O. sativa*
3	14467477	3	13356953	insertion in *O. sativa*
3	14519014	3	13411283	insertion in *O. sativa*
3	14728762	3	13543209	insertion in *O. glaberrima*

3	15651435	3	14362872	insertion in *O. sativa*
3	16059116	3	14827935	insertion in *O. sativa*
3	16155029	3	14922513	insertion in *O. sativa*
3	16857011	3	15661771	insertion in *O. sativa*
3	17050503	3	15820299	insertion in *O. glaberrima*
3	17522139	3	16160194	insertion in *O. sativa*
3	17911432	3	16575497	insertion in *O. glaberrima*
3	20141697	3	19009043	insertion in *O. sativa*
3	21004852	3	19895365	insertion in *O. glaberrima*
3	21147334	3	20033325	insertion in *O. sativa*
3	21198515	3	20084888	insertion in *O. sativa*
3	21228717	3	20116112	insertion in *O. sativa*
3	22848388	3	21603947	insertion in *O. sativa*
3	23158404	3	21856970	insertion in *O. sativa*
3	23277746	3	21956880	insertion in *O. sativa*
3	23470383	3	22125974	insertion in *O. glaberrima*
3	23505898	3	22158886	insertion in *O. sativa*
3	23542933	3	22193450	insertion in *O. sativa*
3	23562711	3	22214793	insertion in *O. glaberrima*
3	23610108	3	22252202	insertion in *O. sativa*
3	24220233	3	22709207	insertion in *O. glaberrima*
3	24498521	3	22935774	insertion in *O. glaberrima*
3	24731186	3	23190007	insertion in *O. sativa*
3	25272800	3	33592728	insertion in *O. glaberrima*
3	25539767	3	23581703	insertion in *O. sativa*
3	25596966	3	23638742	insertion in *O. sativa*
3	25898741	3	23829553	insertion in *O. sativa*
3	26014786	3	23946954	insertion in *O. sativa*
3	26572825	3	24531055	insertion in *O. glaberrima*
3	26698904	3	24647224	insertion in *O. glaberrima*
3	26815101	3	24751979	insertion in *O. glaberrima*
3	26870333	3	24806374	insertion in *O. sativa*
3	27707812	3	25229146	insertion in *O. glaberrima*
3	27778688	3	27778688	insertion in *O. sativa*
3	28130912	3	25635350	insertion in *O. sativa*
3	28261700	3	26033684	insertion in *O. sativa*
3	28641993	3	26403580	insertion in *O. glaberrima*
3	28684915	3	26447023	insertion in *O. glaberrima*
3	28755776	3	26519774	insertion in *O. glaberrima*
3	28829322	3	26593673	insertion in *O. glaberrima*
3	29142858	3	34306612	insertion in *O. sativa*
3	29668203	3	26839837	insertion in *O. sativa*
3	30095692	3	27217619	insertion in *O. sativa*
3	30871123	3	27970290	insertion in *O. sativa*
3	31279236	3	28302618	insertion in *O. glaberrima*
3	31346926	3	28370705	insertion in *O. sativa*
3	31919418	3	28912395	insertion in *O. sativa*
3	31953788	3	28948034	insertion in *O. glaberrima*
3	32042792	3	29037622	insertion in *O. sativa*
3	32081835	3	29071284	insertion in *O. sativa*
3	32615164	3	29533512	insertion in *O. sativa*
3	32657809	3	29575761	insertion in *O. sativa*
3	32685410	7	25338449	insertion in *O. glaberrima*
3	32850769	3	29675472	insertion in *O. sativa*
3	33498009	3	30112785	insertion in *O. glaberrima*
3	34260034	3	30761615	insertion in *O. sativa*
3	34327205	3	30828571	insertion in *O. sativa*
3	34770825	3	31278335	insertion in *O. sativa*
3	35092282	3	31559068	insertion in *O. glaberrima*
3	35436951	3	31885310	insertion in *O. sativa*
3	35732376	3	32150300	insertion in *O. glaberrima*
3	35806887	3	32222454	insertion in *O. sativa*

Supplementary Table 3. Transposition events identified and manually curated in the comparison of the two rice species *O. sativa* and *O. glaberrima*.

Superfamily	Insertions	Excisions
DTH_Harbinger	241	71
DTT_Mariner	137	64
DTM_Mutator	77	20
DTA_hAT	23	1
DTC_CACTA	4	2
Total	482	158

Supplementary Table 4. Average (mean) sequence conservation of promoter and intergenic sequences in different chromosome bins of *O. sativa* and *O. glaberrima*.

Chromsome bin	Promoter	Random	Difference [a]
1	98.22	98.62	28.99%
2	98.03	98.35	19.39%
3	97.8	98.17	20.22%
4	98.06	98.39	20.50%
5	98.33	98.58	17.61%

[a]Difference in sequence divergence between promoter and intergenic sequences

Supplementary Table 5. Wilcoxon rank sum test on comparisons of nucleotide substitutions within rice, barley, wheat, maize and Arabidopsis genes. To normalize for the different sizes of the genes, each gene was divided into 5 equally sized bins and nucleotide substitution frequencies were normalized to substitutions/kb for each bin. Given are the P-values for comparisons of data from all gene bins with all others. P-values smaller than 0.001 were considered significant (marked with *).

Bin pair	Os/Og[a]	Hv/Ta[b]	Maize (IG)[c]	At/Al[d]	Bn(IG)[e]	At/Br[f]	Gm/Pt[g]
1 vs. 2	0.002766	2.2E-16*	4.83E-09*	0.544	0.8738	0.7519	0.9398
1 vs. 3	4.702E-05*	2.2E-16*	5.553E-16*	0.02604	0.3248	0.06229	0.03457
1 vs. 4	0.00543	2.319E-14*	1.93E-08*	0.000138*	0.157	0.1195	0.00727
1 vs. 5	0.696	1.956E-05*	0.04769	1.614E-11*	6.262e-07*	4.733e-06*	1.67e-06
2 vs. 3	0.2863	0.002685	0.02406	0.00453	0.2357	0.1081	0.01868
2 vs. 4	0.7709	0.5643	0.7801	8.609E-06*	0.1153	0.2107	0.00395
2 vs. 5	0.008562	2.2E-16*	0.0002983*	1.518E-13*	7.45e-07*	1.604e-05*	4.75e-07
3 vs. 4	0.1702	0.0003636	0.01026	0.1264	0.6739	0.7326	0.5889
3 vs. 5	0.0002114*	2.2E-16*	6.578E-09*	1.431E-05*	1.219e-04*	0.004051	0.00591
4 vs. 5	0.01644	2.2E-16*	0.0007723*	0.004443	6.224e-04*	0.001096	0.026

[a]Comparison of 442 bi-directional closest homologs from *O. sativa* and *O. glaberrima.*

[b]Comparison of 2,314 bi-directional closest homologs from barley (*H. vulgare*) and wheat (*T. aestivum*)

[c]Comparison of 428 bi-directional closest homeologs within the maize genome that originated from a whole-genome duplication (WGD).

[d]Comparison of 4,133 bi-directional closest homologs from *A. thaliana* and *A. lyrata*.

[e]Comparison of 1,395 bi-directional closest homeologs within the *Brassica napus* genome that originated from a WGD.

[f]Comparison of 536 bi-directional closest homologs from *A. thaliana* and *B. rapa* (the A genome of B. napus)

[g]Comparison of 1,799 bi-directional closest homologs from *Glycine max* and *Populus trichocarpa.*

Supplementary Table 6. Datasets of coding regions (CDS) used for comparative Analyses

Species	genome version	source
Arabidopsis thaliana	9	arabidopsis.org
Arabidopsis lyrata	1.0	genome.jgi-psf.org/Araly1
Brassica napus	5	brassicadb.org/brad
Brassica rapa	1.5	brassicadb.org/brad
Glycine max	1	plantgdb.org/GmGDB
Hordeum vulgare	1.1	pgsb.helmholtz-muenchen.de/plant
Oryza sativa	6	plantgdb.org/OsGDB
Oryza glaberrima	1.0	genome.arizona.edu
Populus trichocarpa	2.2	plantgdb.org/PtGDB
Triticum aestivum	2.2	pgsb.helmholtz-muenchen.de/plant
Zea mays	1.0	maizegdb.org

Chapter 6:

General Discussion

During this PhD project, we studied several aspects of how transposable elements contribute to the evolution of genomes. In the following chapter, I will discuss a few of the main findings.

6.1. Transposable elements are highly active in plants

In Roffler and Wicker (2015) we investigated the activity of DNA transposons by comparison of two closely related rice species, *O. sativa* and *O. glaberrima*. Based on the footprints at the polymorphic TE loci, ee were able to distinguish between insertions and excisions. Based on a subset of 1,821 observations we concluded that approximately 4,000 Class II TEs, or approximately 3.5 % of all DNA TEs, have moved within the last 600,000 years. However, this estimation is veryconservative since it most likely only accounts for the polymorphisms that were fixed in the two species. It has been shown that TE activity can be very tissue specific, such as the dTph1 elements that leads to colored mosaic pattern in the flower of petunia (Gerats *et al.*, 1990), and induced upon various stresses (reviewed by Grandbastien, 1998 and Feschotte *et al.,* 2002). Moreover, we missed those transpositions that caused fitness reduction or even were lethal. Therefore, it is practically certain that the actual number of TE transpositions is still massively underestimated.

The two high quality genomes of *O. sativa* and *O. glaberrima* provided an exceptional opportunity for an analysis of repetitive sequences. However, even here, sequence quality apparently had an influence on the findings. For example, we found more of the shorter and less complex TEs such as *Mariners (DTT)* to be polymorphic in *O. glaberrima* whereas polymorphisms of longer, highly repetitive elements such as the *CACTAs (DTC)* were preferably found in the *O. sativa* sequence (Roffler and Wicker, 2015). To what degree these differences are due to the different levels of activity or abundance in one or the other species remains unclear. However, it appears as if the quality of the sequence and assembly play a role. Therefore, it is still important to produce high-quality genomes, even using “outdated” technologies such as Sanger sequencing, to unravel the full TE content of a species. It will remain challenging to find suitable approaches to investigate TE content based on NGS technology. Possibly long-read sequencing technologies such as PacBio will become sufficiently cheap to be applied to large and complex genomes.

Another of our main findings was that there are significant differences in the ratios of insertions to excisions. We observed that this ratio was bigger for the larger TEs

such as *Mutators (DTM)* and most probably also *CACTA* (too small numbers), meaning that we found less excisions for the larger elements compared to the insertions (Roffler and Wicker, 2015). Additionally, the excision footprints of the more complex elements seemed more severe than those of the smaller elements. Therefore, we hypothesized that elements of the families *Mutator*, *CACTA* and *hAT* are likely to cause either big re-arrangements that we missed in our alignment or that they have a tendency to be so “catastrophic” that they can not be detected anymore (Roffler and Wicker, 2015).

6.2. Non-autonomous elements outnumber their autonomous counterparts

The vast majority of all DNA transposons (at least in grasses) are non-autonomous (Wicker *et al.,* 2010). This phenomenon has been observed for elements from all TE superfamilies. In fact, autonomous DNA TEs are more of an exception. In our work, we fund active families of DNA transposons that show intermediates of the stepwise evolution from autonomous to non-autonomous elements such as the *Mutator* family *DTM_MK* (Roffler and Wicker, 2015) or the *Helitron* family *DHH_Mothra* (Roffler *et al.*, 2015). When considering the dominance of the *Alu* elements in the human genome (Dewannieux *et al.,* 2003), the general principle that non coding TEs outnumber their autonomous counterparts seems also to be true for retrotransposons. Richard Dawkins wrote about the vast amounts of repetitive DNA that “from the point of view of the selfish genes themselves there is no paradox. The true 'purpose' of DNA is to survive, no more and no less. The simplest way to explain the surplus DNA is to suppose that it is a parasite”. Dawkins' perspective proofed right with increasing findings, that much of the repetitive DNA is not only selfish, but that even the originally selfish (but autonomous) TEs have given rise to non-autonomous elements, thereby introducing an even higher level of selfishness.

6.3. It's all double-strand break repair

Retrotransposons, which lack an excision mechanism, have been studied extensively. They contribute significantly to the genome size and thus genome

plasticity (Piegu *et al.,* 2006) and can moreover disrupt genes upon insertion (Miyao *et al.,* 2003). Additionally, the promotor in LTRs has been described to be able to activate nearby genes (Kobayashi *et al.,* 2004). DNA transposons, however, seem to have even stronger effects particularly on the gene content. They have the ability to excise. Excision of TEs creates double-strand-breaks (DSBs) that have to be repaired by the host. Because these were not induced by the host itself, the repair is often error-prone. We showed that excisions often go along with deletions of few bp and in some cases up to several kb. Moreover, we found that excision repair often leads to the insertion of filler sequences (Roffler *et al.,* 2015). Overall, the ratio of excisions to insertions was very low for some DNA transposon superfamilies, suggesting that they cause considerable rearrangements which we would have missed in our approach. Indeed, Wicker *et al.* (2010) showed in a comparative analysis among the three grasses *Brachypodium*, rice and sorhum that genomic fragments of up to 50 kb were duplicated to acceptor sites elsewhere in the genome following TE excisions or insertions.

Additionally, we were able to demonstrate that DNA transposons are preferably located close to genes. Error-prone repair of DSBs caused by TE activity, at least in grasses, induces higher mutation rates in the regulatory regions but also in the coding region of genes. Thereby, especially the excision process seems to have a more severe effect than insertions because of DSB repair (Wicker *et al.,* submitted). Thus TE, but particularly DNA transposons, have an undisputed impact on genome and moreover on gene evolution of grasses.

6.4. Outlook

Our studies have shown that high-quality genome sequences are still needed to appropriately study transposable elements. Such data provides sufficient information to follow TE movement (given that the compared organisms did not diverge too long ago).

Most of our findings on TE activity are based on the initial alignment of the genomes of *O. sativa* and *O. glaberrima* and focused on small, non-autonomous elements. Because the alignment consists of overlapping, highly conserved

fragments of 5 kb, we are aware that the investigated gap size is thus restricted to these 5 kb. Moreover, we only considered highly homologous regions, which would exclude for more drastic re-arrangements.

As the genome of *O. glaberrima* (Wang *et al.,* 2014) was just the first of ten high-quality rice genomes that are currently sequenced, this data would now allow an overall comparison between the TE content of all ten species. The fact that they all diverged within the past 15 million years would allow highly detailed comparisons for at least some genomic regions and a very broad assessment of TE activity. The additional information from the eight additional genome sequences could be used to infer many intermediate states and thus to reconstruct the genomes' history and the possible implications of TEs.

6.5. References

Dawkins R: **The Selfish Gene.** *Oxford University Press 1976*.

Dewannieux M, Esnault C, Heidmann T: **LINE-mediated retrotransposition of marked Alu sequences.** *Nature Genet. 2003,* **35:**41–48.

Feschotte C, Jiang N, Wessler SR: **Plant transposable elements: where genetics meets genomics.** *Nature Reviews Genetics 2002,* **(3):**329-341.

Gerats AG, Huirs H, Vrijlandt E, Maraña C, Souer E, Beld M: **Molecular characterization of a nonautonomous transposable element (dTph1) of petunia.** *Plant Cell 1990,* **2(11):**1121-8.

Grandbastien MA: **Activation of plant retrotransposons under stress conditions.** *Trends in Plant Science 1998,* **3(5):**181-187.

Kobayashi S, Goto-Yamamoto N, Hirochika H: **Retrotransposon-induced mutations in grape skin color.** *Science 2004,* **304(5673):**982.

Miyao A, Tanaka K, Murata K, *et al.*: **Target Site Specificity of the *Tos17* Retrotransposon Shows a Preference for Insertion within Genes and against Insertion in Retrotransposon-Rich Regions of the Genome.** *Plant Cell 2003,* **15(8):**1771-1780.

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H *et al.*: **Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice.** *Genome Res. 2006,* **16:**1262–9.

Roffler S and Wicker T: **Genome-wide comparison of Asian and African rice reveals high recent activity of DNA transposons.** *Mobile DNA 2015,* **6:**8.

Roffler S, Menardo F, Wicker T: **The making of a genomic parasite – the *Mothra* family sheds light on the evolution of *Helitrons* in plants.** *Mobile DNA 2015,* **6:**23.

Wang M *et al.*: **The genome sequence of African rice (Oryza glaberrima) and evidence for independent domestication.** *Nature Genetics 2014*, **46(9)**:982-988.

Wicker T, Buchmann JP, Keller B: **Patching gaps in plant genomes results in gene movement and erosion of colinearity.** *Genome Res. 2010,* **20:**1229-1237.

Wicker T, Yu Y, Haberer G, Mayer KFX, Reddy Marri P, Rounsley S, Chen M, Zuccolo A, Panaud O, Wing RA, Roffler S: **DNA transposons specifically accelerate evolution of genes in rice and other grasses.** *Submitted 2016*.

7. Acknowledgments

First of all I would like to thank Beat Keller and Thomas Wicker for having me in their group and giving me the possibility to do such work. Moreover, I would like to thank Christian von Mehring for being part of my PhD committee. It would be an endless list to name all my friends and former and present co-workers here. I would like to thank all of you for the good times we spent in the lab, on the roof or in the bunker. Finally, it is my family, in especial my parents and Miri who supported me throughout all ups and downs.